Test a (causal) hypothesis with observational data (THE III)

Research Design and Methods in Quantitative Research - Fall 2024

Álvaro Canalejo-Molero

2024-11-15

Instructions

Please read and follow the guidelines below carefully. Then, complete the exercises and report the results in a Quarto document. Compile the Quarto document in PDF and submit both the compiled PDF and `.qmd` files within the deadline.

Further instructions about the submission are below.

Preparation step 1: Install R and RStudio

To complete this exercise, you will need **R** and **RStudio** again. If you have already downloaded and installed R and RStudio, you can skip this part. Otherwise, please, download and install them from:

- R
- RStudio

A tutorial on how to start using R and R Studio is here. ***Please contact the tutor and collaborate with your classmates in case of doubts or if you need any help.***

Preparation step 2: Prepare a Quarto Document

Open RStudio, create a new Quarto document (`.qmd`), and set the output format to PDF. Make sure your Quarto installation is up-to-date:

```
# Install Quarto if needed
## Run this line in a separate script or the Quarto document will not compile
# install.packages("quarto")
```

You can find help on how to set up a Quarto document here.

Preparation step 3: Download the CSES Integrated Module Dataset (IMD)

The Comparative Study of Electoral Systems (CSES) is a standardized post-electoral cross-national survey that covers most democratic countries worldwide. It has been running since 1996 until now. The CSES Integrated Module Dataset (IMD) include selected variables from CSES Modules 1 through 5 into a single, harmonized longitudinal dataset. Variables included in the IMD must appear in at least three CSES Modules, up to and including Module 5. The dataset encompasses all polities that participated in these modules, featuring over 395,000 individual-level observations from 230 elections across 59 polities. Notable features of the CSES IMD include harmonized numerical codes for parties and coalitions within polities across modules, a pre-coded variable for incumbent vote choice, political information indexes, comprehensive macro-level data spanning over 25 years, and detailed within-dataset labels for all parties and coalitions.

For this exercise, you will need to register in the CSES webpage and download the CSES IMD. For easy integration with R, I recommend downloading the following file:

- cses_imd_r.zip

You will also need the codebook.

Exercises

Exercise 1: Draw a DAG of the causal hypothesis

You will use the CSES IMD data to test one of the most prominent theories in political science: the economic voting theory. This theory states that citizens decide their vote based on the state of the economy, so that they are more likely to support the incumbent (i.e., the outgoing governing party or candidate) if the economy works well and less likely is the economic situation is bad (for a review, see Lewis-Beck and Stegmaier, 2018). Focusing on citizens' perceptions rather than objective economic indicators, the following (causal) hypothesis follows:

> **Hypothesis**: Negative (positive) economic evaluations reduce (increase) the probability of voting for the incumbent.

The CSES data contains one variable that we can use as our independent ("treatment") variable, as it is operationalized as follows:

- `IMD3013_1`: Would you say that over the past twelve months, the state of the economy in [COUNTRY] has gotten better, stayed about the same, or gotten worse?

It also contains a variable that we can use to measure voting for the incumbent (i.e., our dependent variable):

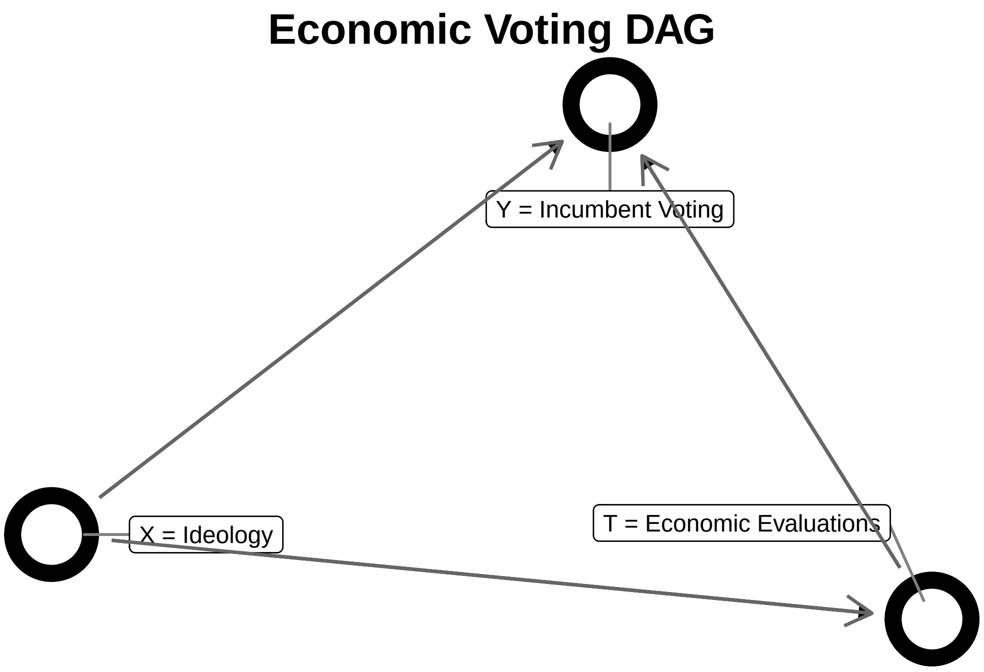
- `IMD3002_OUTGOV`: Whether or not the respondent cast a ballot for the outgoing incumbent.

Since we want to approximate a causal test, we will need to control for some variables in order to rule out endogeneity concerns. However, we do not want to do this blindly but based on theoretical reasoning. As you will see in the codebook, there are many variables that we can use (and more that we can construct) to try to isolate the causal relationship between economic evaluations and voting for the incumbent, but we do not want to use them all. To guide our model specification, **please draw a DAG of the theoretical relationship at hand**. This will guide our next choices, so it is important that it is done carefully.

You can draw a DAG in R using the packages `dagitty` and `ggdag` (see installation and loading below). Alternatively, you can draw it with any other program or by hand and upload it to the .qmd document as an image. Please comment your decisions.

PS: Remember to simplify and do not include every variable you think it could be involved in the relationship, but only the most important. Also, group variables under broader concepts to avoid overfitting the DAG (e.g., 'socio-economic conditions' instead of 'employment status' and 'income').

```r
# Install packages
#install.packages(c("dagitty", "ggdag"))

# Load necessary libraries
library(tidyverse)
library(dagitty)
library(ggdag)


# Define the DAG
dag <- dagitty('
dag {
  Economic_Evaluations [exposure]
  Incumbent_Voting [outcome]
  Ideology [confounder]
  Economic_Evaluations -> Incumbent_Voting
  Ideology -> Economic_Evaluations
  Ideology -> Incumbent_Voting
}
')

# Convert DAG to tidy format
tidy_dag <- tidy_dagitty(dag)

# Add labels for clarity
tidy_dag <- tidy_dag %>%
  mutate(label = case_when(
    name == "Economic_Evaluations" ~ "T = Economic Evaluations",
    name == "Incumbent_Voting" ~ "Y = Incumbent Voting",
    name == "Ideology" ~ "X = Ideology",
    TRUE ~ name
  ))

# Visualize the DAG
ggdag(tidy_dag, text = FALSE) +
  geom_dag_point(size = 10, color = "white") + # Cool blue points
  geom_dag_label_repel(aes(label = label), box.padding = 3,
                       size = 3) + # Labels with repulsion
  geom_dag_edges_link(arrow = grid::arrow(length = unit(0.4, "cm")),
                      edge_colour = "gray40") + # Gray edges with arrows
  theme_void() + # Remove background
  labs(
    title = "Economic Voting DAG"
  ) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
    plot.caption = element_text(hjust = 0.5, size = 10)
  )
```



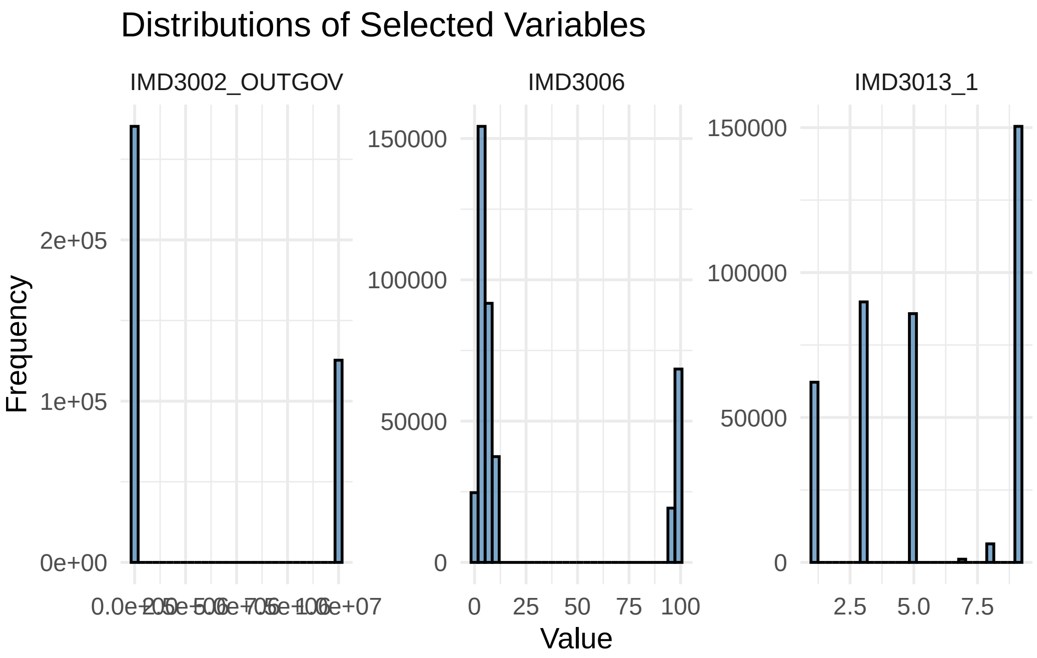
Exercise 2: Prepare your data

Based on your DAG, declare what variables you will use for the analysis. Please inspect and transform them when necessary (e.g., recode missing values, inverse scales, etc.). Report your code and comment your decisions.

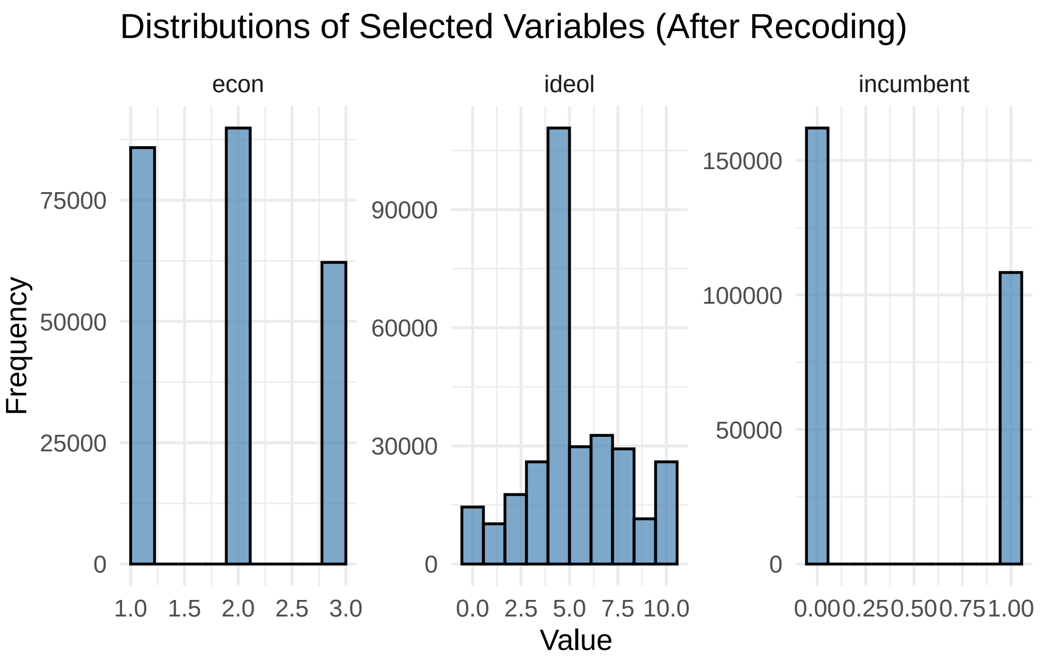
```
# Load the data
load("C:/Users/acana/Dropbox/Research/GitHub/teaching/rdmqr_unilu2024/00_take_home_exercises/

# Inspect the summary of the variables
variables_to_inspect <- c("IMD3002_OUTGOV", "IMD3013_1", "IMD3006")
summary_stats <- cses_imd %>%
  select(all_of(variables_to_inspect)) %>%
  summary()

print(summary_stats)
```

```
 IMD3002_OUTGOV      IMD3013_1        IMD3006
 Min.   :      0   Min.   :1.000   Min.   : 0.00
 1st Qu.:      0   1st Qu.:3.000   1st Qu.: 5.00
 Median :      1   Median :5.000   Median : 6.00
 Mean   :3168164   Mean   :5.492   Mean   :25.85
 3rd Qu.:9999996   3rd Qu.:9.000   3rd Qu.:10.00
 Max.   :9999999   Max.   :9.000   Max.   :99.00
```

```r
# Visualize distributions
cses_imd %>%
  select(all_of(variables_to_inspect)) %>%
  gather(key = "Variable", value = "Value") %>%
  filter(!is.na(Value)) %>%
  ggplot(aes(x = Value)) +
  geom_histogram(bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +
  facet_wrap(~ Variable, scales = "free") +
  theme_minimal() +
  labs(title = "Distributions of Selected Variables",
       x = "Value",
       y = "Frequency")
```



```r
# Recode IMD3006: Set values > 10 to NA
cses_imd <- cses_imd %>%
  mutate(ideol = ifelse(IMD3006 > 10, NA, IMD3006))

# Recode IMD3013_1: Set values > 5 to NA, and recode to make higher values better
cses_imd <- cses_imd %>%
  mutate(
    econ = NA, # Set invalid values to NA
    econ = ifelse(IMD3013_1 == 1, 3, econ), # Gotten Better -> 3
    econ = ifelse(IMD3013_1 == 3, 2, econ), # Stayed the Same -> 2
    econ = ifelse(IMD3013_1 == 5, 1, econ)  # Gotten Worse -> 1
  )

# Recode IMD3002_OUTGOV: Set values > 1 to NA
cses_imd <- cses_imd %>%
  mutate(incumbent = ifelse(IMD3002_OUTGOV > 1, NA, IMD3002_OUTGOV))

# Variables to inspect
variables_to_inspect <- c("ideol", "econ", "incumbent")

# Replot the distributions
cses_imd %>%
  select(all_of(variables_to_inspect)) %>%
  gather(key = "Variable", value = "Value") %>%
  filter(!is.na(Value)) %>%
  ggplot(aes(x = Value)) +
  geom_histogram(bins = 10, fill = "steelblue", color = "black", alpha = 0.7) +
  facet_wrap(~ Variable, scales = "free") +
  theme_minimal() +
  labs(title = "Distributions of Selected Variables (After Recoding)",
       x = "Value",
       y = "Frequency")
```



Exercise 3: Test your hypothesis with an ordinary-least squares (OLS) multiple regression model

Run an OLS regression model to test your hypothesis. Use the function `lm()` for that, or you can use more complicated functions if preferred. Then comment on your choices and your results. Based on them, does the evidence support your hypothesis?

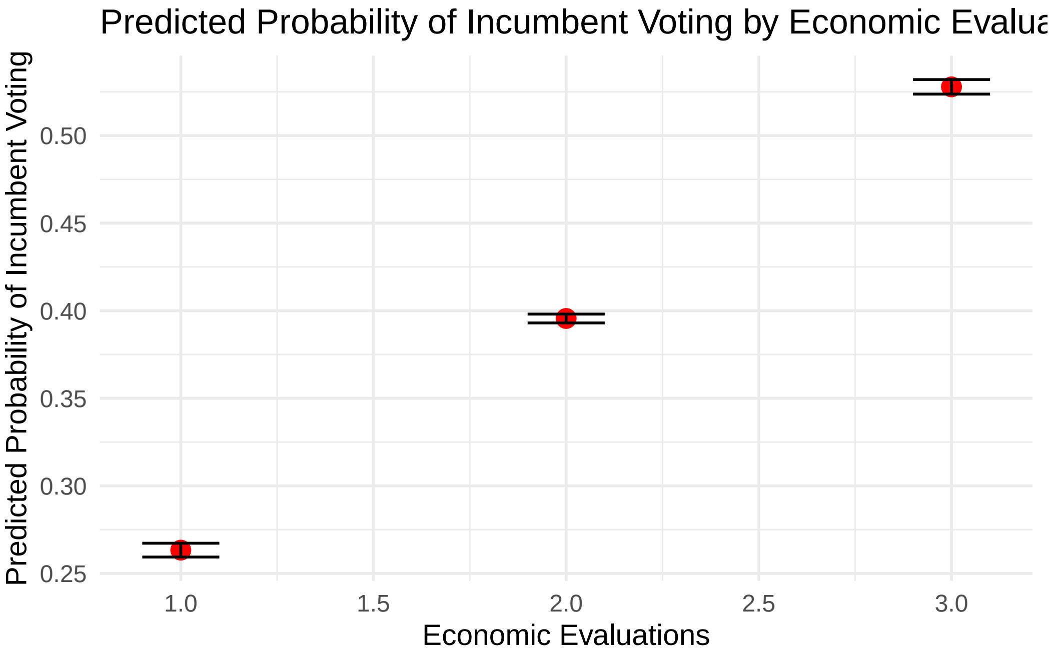
Optional: you can plot the predicted probabilities of voting for the incumbent based on economic evaluations. This may be helpful for interpreting your results.

```r
# Run OLS regression
model <- lm(incumbent ~ econ + ideol, data = cses_imd)

# Display summary of the model
summary(model)
```

```
Call:
lm(formula = incumbent ~ econ + ideol, data = cses_imd)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-0.5914 -0.3906 -0.2583  0.5183  0.8103

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.057463   0.004259   13.49   <2e-16 ***
econ        0.132261   0.001612   82.06   <2e-16 ***
ideol       0.013713   0.000487   28.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4747 on 136622 degrees of freedom
  (259172 observations deleted due to missingness)
Multiple R-squared:  0.05262,   Adjusted R-squared:  0.05261
F-statistic:  3794 on 2 and 136622 DF,  p-value: < 2.2e-16
```

```r
# Generate predicted values with confidence intervals for specific levels of econ
predicted_data <- data.frame(
  econ = c(1, 2, 3),
  ideol = mean(cses_imd$ideol, na.rm = TRUE) # Set ideol to its mean
)

# Add predictions and confidence intervals
predictions <- predict(model, newdata = predicted_data, interval = "confidence")
predicted_data <- cbind(predicted_data, predictions)

# Plot the predicted probabilities with error bars
ggplot(predicted_data, aes(x = econ, y = fit)) +
  geom_point(size = 3, color = "red") + # Predicted points
  geom_errorbar(aes(ymin = lwr, ymax = upr), width = 0.2, color = "black") + # Error bars
  theme_minimal() +
  labs(
    title = "Predicted Probability of Incumbent Voting by Economic Evaluations",
    x = "Economic Evaluations",
    y = "Predicted Probability of Incumbent Voting"
  )
```



Exercise 4 (additional): Run an additional analysis of your choice

This exercise is not mandatory, but it serves only to opt for the maximum grade (6).

Is there any other statistical test you could run to further support or disprove the hypothesis? Please think on the observable implications of the theory that could be tested with the CSES IMD data and provide an additional test. It can be either another regression specification or a different statistical analysis. Finally, comment on your decisions and results, and discuss them together with the results of the previous exercise. You can be as creative as you want here; it is the final exercise, so ***enjoy yourself***!

Bibliography

Lewis-Beck, M. S., & Stegmaier, M. (2018). Economic voting. The Oxford handbook of public choice, 1, 247-265.