Take Home Exercise 3

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

Nicolas Waser

2024-12-12

```
options(repos = c(CRAN = "https://cloud.r-project.org")) # set CRAN mirror
```

Disclaimer! For this exercise I made use of the following AI Generation tools: Deepl.com for translation purposes, the integrated Copilot tool in R Studio for generating code and text and the premium version of OpenAI's GPT-4o and o1, as well as the free version of Anthropics Claude 3.5 Sonnet for generating code and assisting me in solving the tasks.

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
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
install.packages(c("dagitty", "ggdag", 'lfe', 'ggpubr', 'mice', 'car', 'stargazer'))
```

```
The downloaded binary packages are in
    /var/folders/3k/4r5z06ts7bl41nkd_6d0dcyh0000gn/T//Rtmpr9xggQ/downloaded_packages
```

```r
library(dagitty)
library(ggdag)
```

```
Attaching package: 'ggdag'

The following object is masked from 'package:stats':

    filter
```

```r
library(dplyr)
library(ggplot2)
library(ggpubr)
library(lfe)
```

```
Loading required package: Matrix

Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack

Registered S3 method overwritten by 'lfe':
  method    from
  nobs.felm broom
```

```r
library(mice)
```

```


Attaching package: 'mice'

The following object is masked from 'package:stats':

    filter

The following objects are masked from 'package:base':

    cbind, rbind
```

```r
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode
```

```
The following object is masked from 'package:purrr':

    some
```

```
library(stargazer)
```

```
Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
rm(list = ls())
#cses_imd <- load('/Users/nicolaswaser/New-project-GitHub-first/R/Take-Home-Ex.-3/Input Data/
library(readr)
cses_imd <- read_csv("cses_imd.csv")
```

```
New names:
Rows: 395797 Columns: 407
-- Column specification
-------------------------------------------------------- Delimiter: "," chr
(11): IMD1001, IMD1002_VER, IMD1002_DOI, IMD1004, IMD1005, IMD1006, IM... dbl
(395): ...1, IMD1003, IMD1006_UN, IMD1006_REG, IMD1006_OECD, IMD1006_EU... date
(1): IMD1011_1
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.
* `` -> `...1`
```

Independent Variable (IV): IMD3013_1 -> STATE OF ECONOMY (OVER PAST 12 MONTHS)

"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?"

Categorical: 1 = Better, 2 = Same, 3 = Worse, (7 = refused, 8 = don't know, 9 = missing)

Dependent Variable (DV): IMD3002_OUTGOV -> VOTE CHOICE: CURRENT MAIN ELECTION - VOTE FOR OUTGOING GOVERNMENT (INCUMBENT)

Whether or not the respondent cast a ballot for the outgoing incumbent.

Binary: 0 = Did not vote for incumbent, 1 = Voted for incumbent, (999999_ for other answers)

Selecting possible Control Variables: Socio-demographic Characteristics: Age, IMD2001_1 (age in years, continuous, 015-115), IMD2001_2 (age in categories, 0-6): The older, maybe more likely to vote for the incumbent and more likely to be satisfied with economy??? Gender, IMD2002 (0,1): Female maybe more likely to vote for the incumbent??? More satisfied with economy??? Urban or Rural Residence, IMD2007 (1-4): Urban residents more likely to be satisfied with economy? Education, IMD2003 (0-6): The better educated, the more likely to be satisfied with economy! Maybe more likely to vote for the incumbent? Household Income, IMD2006 (1-5): The richer, the more likely to be satisfied with economy! Maybe more likely to vote for the incumbent? Employment Status, IMD2014 (0-10): Employed more likely to be satisfied with economy and more likely to vote for the incumbent!

Political/Partisan Predispositions: Party Identification/Partisan, IMD3005_1 (0,1), “Do you feel close to any one party?”, (IMD3005_2): If yes, the more likely to vote for own party, regardless of incumbent/challenger status and economy? Ideology, IMD3006 (left-right self-placement, 0-10): Useful for Party id?

Political Information/Politically Informed: IMD3015_1 -> POLITICAL INFORMATION: DICHOTOMIZED ITEM - 1ST IMD3015_2 -> POLITICAL INFORMATION: DICHOTOMIZED ITEM - 2ND IMD3015_3 -> POLITICAL INFORMATION: DICHOTOMIZED ITEM - 3RD IMD3015_4 -> POLITICAL INFORMATION: DICHOTOMIZED ITEM - 4TH 0 = Incorrect, 1 = Correct

IMD3015_A -> POLITICAL INFORMATION: SCALE - CSES MODULE 1 (0-3 SCALE) IMD3015_B -> POLITICAL INFORMATION: SCALE - CSES MODULE 2 (0-3 SCALE) IMD3015_C -> POLITICAL INFORMATION: SCALE - CSES MODULE 3 (0-3 SCALE) IMD3015_D -> POLITICAL INFORMATION: SCALE - CSES MODULE 4 (0-4 SCALE) 0-3, 0-4 number of correct answers If more informed, the less likely, that the state of economy does matter for voting for incumbent!

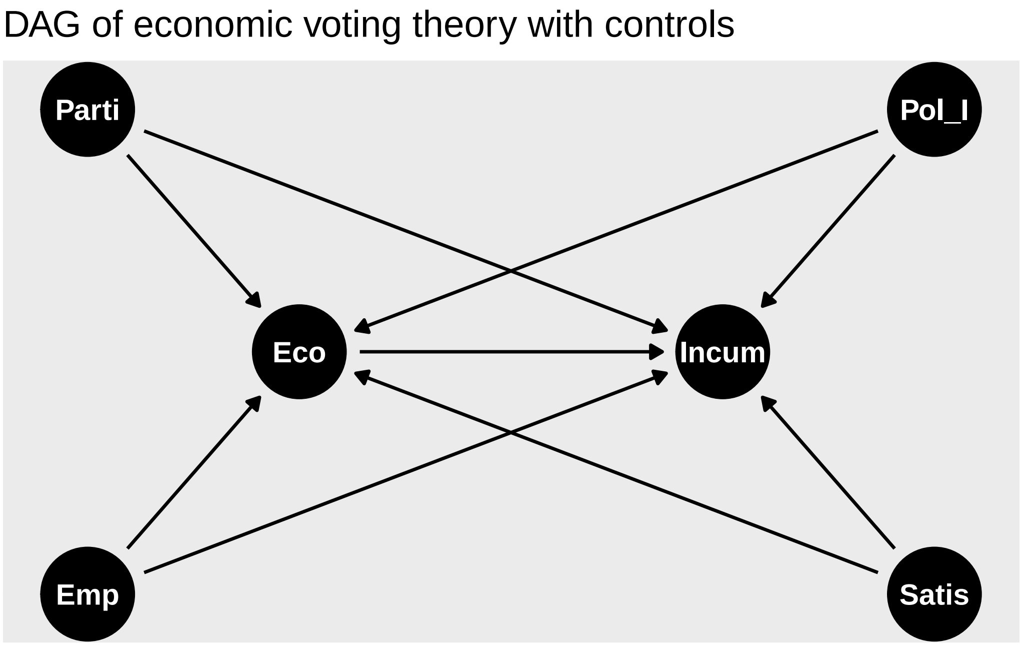
Satisfaction with Democracy: IMD3010 -> SATISFACTION WITH DEMOCRACY (1-6) The more satisfied with Dem., the less likely the state of Economy has an effect on voting for incumbent!

IMD3011 -> EFFICACY: WHO IS IN POWER CAN MAKE A DIFFERENCE (1-5) The more agreement, with this statement, the more likely the state of economy has an effect on voting for incumbent? IMD3012 -> EFFICACY: WHO PEOPLE VOTE FOR MAKES A DIFFERENCE (1-5) The more agreement, with this statement, the more likely the state of economy has an effect on voting for incumbent?

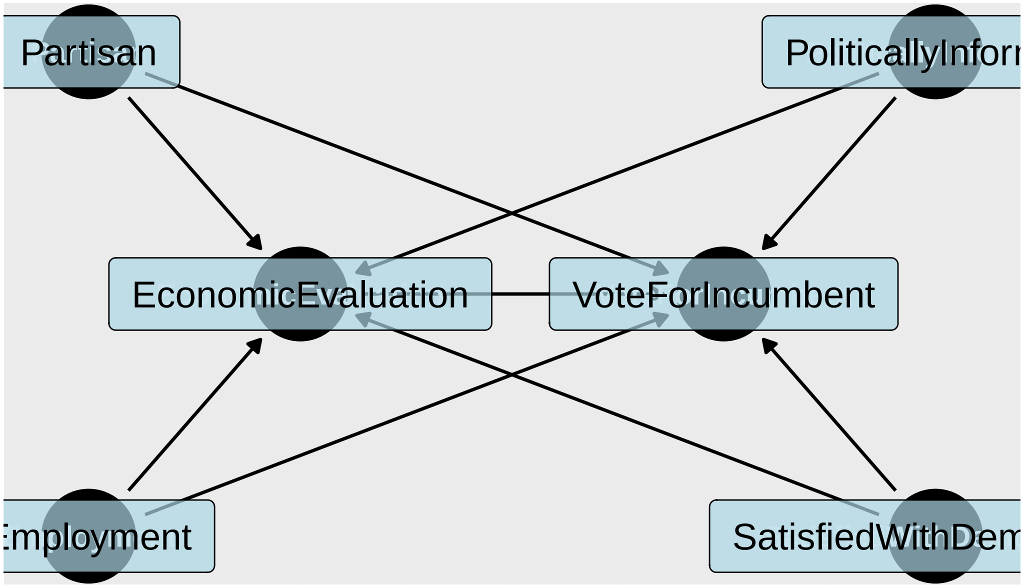
I pick the following Control Variables: Employment Status (Emp), IMD2014 (0-10): I strongly suspect that employed people are more likely to vote for the incumbent and more likely to be satisfied with the economy. Unemployed people, on the other hand, might be tempted to vote for the challenger, hoping for a change in the economy and an improvement in their personal situation. Partisan (Parti), IMD3005_1 (0,1): If the respondent feels close to any party, they are more likely to vote for their party, regardless of the incumbent/challenger status and the state of the economy. Individuals with strong partisan or ideological leanings may tend to view economic conditions through a partisan lens, shaping both their perceptions of the economy and their vote choices. However, as will be shown later, I was unable to construct the variable in such a way that it would exactly represent my theory. Nevertheless, it is worth including it in the model even in its simplified form. Political Information (Pol_I), IMD3015_A (0-3), IMD3015_B (0-3), IMD3015_C (0-3), IMD3015_D (0-4): The more politically informed a respondent is, the less likely the state of the economy should have an effect on voting for the incumbent. People who are more informed may have more nuanced perceptions of the economy and be more aware of the incumbent's responsibility for current conditions. Satisfaction with Democracy (Satis), IMD3010 (1-6): The more satisfied with democracy, the less likely the state of the economy should have an effect on voting for the incumbent.If someone is already dissatisfied with the incumbent's handling of other policy areas or has low trust, they might be more inclined to interpret economic signals negatively and vote against the incumbent.

The DAG will be as follows: IV (Eco) (IMD3013_1) -> DV (Incum) (IMD3002_OUTGOV) Control Variables -> DV Control Variables -> IV

```
# Simple DAG
dag_text <- "dag {
  Eco [exposure,pos=\"-0.5,0\"]
  Incum [outcome,pos=\"0.5,0\"]
  Emp [pos=\"-1,-1\"]
  Parti [pos=\"-1,1\"]
  Pol_I [pos=\"1,1\"]
  Satis [pos=\"1,-1\"]
  Eco -> Incum
  Emp -> Eco
  Parti -> Eco
  Pol_I -> Eco
  Satis -> Eco
  Emp -> Incum
  Parti -> Incum
  Pol_I -> Incum
  Satis -> Incum}"
g <- dagitty(dag_text)
ggdag(g, layout = "auto") + theme_dag_grey() + theme(legend.position = "right") + ggtitle("DA
```



```
# Labeled DAG
dag_full_text <- "dag {
  EconomicEvaluation [exposure,pos=\"-0.5,0\"]
  VoteForIncumbent [outcome,pos=\"0.5,0\"]
  Employment [pos=\"-1,-1\"]
  Partisan [pos=\"-1,1\"]
  PoliticallyInformed [pos=\"1,1\"]
  SatisfiedWithDemocracy [pos=\"1,-1\"]
  EconomicEvaluation -> VoteForIncumbent
  Employment -> EconomicEvaluation
  Partisan -> EconomicEvaluation
  PoliticallyInformed -> EconomicEvaluation
  SatisfiedWithDemocracy -> EconomicEvaluation
  Employment -> VoteForIncumbent
  Partisan -> VoteForIncumbent
  PoliticallyInformed -> VoteForIncumbent
  SatisfiedWithDemocracy -> VoteForIncumbent}"
g_full <- dagitty(dag_full_text)
ggdag(g_full, layout = "auto") + geom_dag_label(aes(label = name),
                 size = 5, label.padding = unit(0.5,   "lines"),
                 label.r = unit(0.15, "lines"), alpha = 0.7,
                 fill = "lightblue", color = "black"   ) +
  theme_dag_gray() + ggtitle("DAG of economic voting theory with controls")
```


DAG of economic voting theory with controls

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.

Operationalization of the variables: IV, (IMD3013_1): Economic Evaluation (1-3, 7-9 -> NA) DV, (IMD3002_OUTGOV): Vote for Incumbent (0,1, 999999_ -> NA) Control Variables:

Employment, (IMD2014) (0-10; 0-5 in Labor Force, 6-10 not in Labor Force, 11-12, 97-99 -> NA) Problem: Half in Labor Force, half not in Labor Force, how to operationalize? Mark 1-5 as most to least employed?, Or: 0-3 = employed (1), 4-5 = unemployed (0), drop 6-10?

Partisan, (IMD3005_1) (0 = No, 1 = Yes, 7-9 -> NA) Problem: Ideally I would create a new variable: ((Partisan = Yes) & (Incumbent = same Party)). But I would need additional information to create this control variable: The Incumbent Party & the respondent's party. This information is either not contained in the data set or it is, but in such a way, that it is unusable. From a theoretic point of view it would be the most promising control variable, as it makes intuitively sense, that being partisan would make you care less about the state of the economy and more about the party colors of the incumbent. If he is from your party, you are more likely to vote for him and vice versa, regardless of the state of the economy.

```
# Partisan Control Variable Construction Example ----------------------------------
#df3 <- cses_imd[, c("IMD3013_1", "IMD3002_OUTGOV", "IMD2014", "IMD3005_1",
   #                 "IMD3005_3", INCUMBENT_PARTY_CODE, "IMD3015_A", "IMD3015_B",
    #                "IMD3015_C", "IMD3015_D", "IMD3010")]
#colnames(df3) <- c("Eco_Eval", "Vote_For_Incu", "Employment", "Partisan",
 #                  "Resp_Party", "Incum_Party", "Pol_Info_A", "Pol_Info_B",
   #                "Pol_Info_C", "Pol_Info_D", "Sat_with_Dem")
#df3 <- df3 %>% mutate( # Creating partisan control variable
 #   political_predisposition = case_when(
      # Respondent identifies with incumbent's party
  #    Vote_For_Incu == 1 & IMD3005_3 == INCUMBENT_PARTY_CODE ~ "Incumbent",
      # Respondent identifies with a different party
   #   Vote_For_Incu == 1 & IMD3005_3 != INCUMBENT_PARTY_CODE ~ "Opposition",
      # Respondent does not identify with any party
    #  Vote_For_Incu == 0 ~ "Non-partisan",
      # If no data or missing
     # TRUE ~ NA_character_))     # INCUMBENT_PARTY_CODE is a Placeholder
```

This approach is not feasible, but I will include the Partisan variable as a control variable nonetheless, as it is the best available proxy for the theoretical concept of being partisan and I suspect it will still have a significant effect on the final regression

Political Information, IMD3015_A (0-3), IMD3015_B (0-3), IMD3015_C (0-3), IMD3015_D (0-4), (0-3/0-4 number of correct answers, 9 -> NA) Problem: How to operationalize? Sum or aggregate all 4 variables? Or pick one?

Satisfaction with Democracy, IMD3010 (1-5, 6 = neutral, 6 -> 3, 7-9 -> NA) Problem: 6 = neutral -> Recode to 3!

Checking Data Structure:

```
df <- cses_imd[, c("IMD3013_1", "IMD3002_OUTGOV", "IMD2014", "IMD3005_1",
                   "IMD3015_A", "IMD3015_B", "IMD3015_C", "IMD3015_D",
                   "IMD3010")] # New df with the variables of interest
colnames(df) <- c("Eco_Eval", "Vote_For_Incu", "Employment", "Partisan",
                  "Pol_Info_A", "Pol_Info_B", "Pol_Info_C", "Pol_Info_D",
                  "Sat_with_Dem") # renaming columns
str(df) # Checking the structure of the new df
```

```
tibble [395,797 x 9] (S3: tbl_df/tbl/data.frame)
 $ Eco_Eval     : num [1:395797] 9 9 9 9 9 9 9 9 9 9 ...
 $ Vote_For_Incu: num [1:395797] 0e+00 0e+00 0e+00 1e+07 0e+00 ...
 $ Employment   : num [1:395797] 6 8 10 8 8 1 7 7 7 1 ...
 $ Partisan     : num [1:395797] 0 0 0 1 0 1 1 0 0 1 ...
 $ Pol_Info_A   : num [1:395797] 9 9 9 9 9 9 9 9 9 9 ...
 $ Pol_Info_B   : num [1:395797] 2 0 2 1 0 1 1 2 3 1 ...
 $ Pol_Info_C   : num [1:395797] 9 9 9 9 9 9 9 9 9 9 ...
 $ Pol_Info_D   : num [1:395797] 9 9 9 9 9 9 9 9 9 9 ...
 $ Sat_with_Dem : num [1:395797] 4 2 4 2 5 5 5 4 4 2 ...
```

All values are numerical values, no factors, no characters, no NAs

```r
#table(df$Eco_Eval) ## Economic Evaluation
prop.table(table(df$Eco_Eval))*100 # Show distribution in %
```

```
         1          3          5          7          8          9 
15.7065364 22.7050736 21.6853589  0.2786782  1.6169905 38.0073624 
```

About 40% invalid answers -> mark as NA -> impute or remove? Simply removing could lead to a substantial loss of data, which may compromise the statistical power and generalizability of results. Imputation is possible, predictive mean matching (pmm) is probably best suited for this task.

```r
#table(df$Vote_For_Incu) ## Vote for Incumbent
prop.table(table(df$Vote_For_Incu))*100
```

```
         0          1    9999996    9999997    9999998    9999999 
40.9422002 27.3761549 10.9154440  3.3082616  0.7056648 16.7522745 
```

About 30% invalid answers-> mark as NA -> impute or remove? Again simply removing could lead to a substantial loss of data and having missing values in the dependent variable is more problematic in a regression model, because these rows cannot directly contribute to the regression analysis. Therefore, it is better to remove the rows with missing values in the dependent variable altogether (listwise deletion). Note: Having this many missing values in the dependent and independent variable is not at all ideal. In a different task setting, it might make sense to look around for alternatives.

```r
#table(df$Partisan) ## Partisan
prop.table(table(df$Partisan))*100
```

```
        0         1         7         8         9 
50.097651 42.291629  1.500517  2.249385  3.860818 
```

```
#table(df$Sat_with_Dem) ## Satisfaction with Democracy
prop.table(table(df$Sat_with_Dem))*100
```

```
         1          2          4          5          6          7          8 
10.0574284 44.9457677 27.3051590 10.8901786  0.9209266  0.7056648  3.1379722 
         9 
 2.0369028 
```

Both about 6% invalid answers -> mark as NA -> Imputation possible! In order to not further reduce the sample size, I will impute the missing values, which in this case should be easily doable. Regular Median Imputation should suffice. 'Sat_with_Dem' needs to be recoded, as 6 is a neutral answer ('neither safisfied or dissatisfied') and should be squarely in the middle of the scale from 1-5. Therefore it will be recoded to 3.

```
#table(df$Pol_Info_A) ## Politically Informed
prop.table(table(df$Pol_Info_A))*100 # 90% invalid -> remove!
```

```
        0         1         2         3         9 
 1.253926  2.595522  2.890623  2.304212 90.955717 
```

```
# table(df$Pol_Info_B) -------------------------------------------------------------
#prop.table(table(df$Pol_Info_B))*100 # 86% invalid -> remove!
#table(df$Pol_Info_C)
#prop.table(table(df$Pol_Info_C))*100 # 82% invalid -> remove!
#table(df$Pol_Info_D)
#prop.table(table(df$Pol_Info_D))*100 # 84% invalid -> remove!
```

All politically informed variables have more than 80% invalid answers. There is no way to impute these values, removing this much from the sample is also not reasonable, so I will remove them from the dataset. Let's try the other politically informed variables that are binary, maybe this will lead to better results:

```
df <- cses_imd[, c("IMD3013_1", "IMD3002_OUTGOV", "IMD2014", "IMD3005_1",
                   "IMD3015_1", "IMD3015_2", "IMD3015_3", "IMD3015_4",
                   "IMD3010")]
colnames(df) <- c("Eco_Eval", "Vote_For_Incu", "Employment", "Partisan",
                  "Pol_Info_1", "Pol_Info_2", "Pol_Info_3", "Pol_Info_4",
                  "Sat_with_Dem")
```

```
#table(df$Pol_Info_1)
prop.table(table(df$Pol_Info_1))*100 # 50% invalid -> remove!
```

```
        0         1         7         8         9 
 8.605674 40.805514  0.361549 10.131456 40.095807 
```

```
# table(df$Pol_Info_2) -------------------------------------------------------------
#prop.table(table(df$Pol_Info_2))*100 # 54% invalid -> remove!
#table(df$Pol_Info_3)
#prop.table(table(df$Pol_Info_3))*100 # 56% invalid -> remove!
#table(df$Pol_Info_4)
#prop.table(table(df$Pol_Info_4))*100 # 90% invalid -> remove!
```

Still more than 50% invalid answers for all 4 subsets, so I have to cut this control variable altogether. It does not make sense including it.

```
df<-cses_imd[, c("IMD3013_1","IMD3002_OUTGOV","IMD2014","IMD3005_1","IMD3010")]
colnames(df)<-c("Eco_Eval","Vote_For_Incu","Employment","Partisan","Sat_with_Dem")
```

```
#table(df$Employment) ## Employment
prop.table(table(df$Employment))*100
```

```
          0           1           2           3           4           5 
 5.05562195 27.23997403  4.66097520  0.94164433  0.49570866  4.11827275 
          6           7           8           9          10          11 
 3.74030122 13.33006566  6.49878599  1.19404644  1.33048002  0.02880264 
         12          97          98          99 
 0.02400220  0.31379722  0.12708535 30.90043633 
```

About 30% invalid answers-> mark as NA -> impute or remove? Kill variable? With this many missing values, it would probably be best to remove the column from the df. But I already killed another control variable, so I will keep it here and attempt to impute the missing values with the 'pmm' method from the mice package. Another problem is the coding of the variable. The variable is coded in a way that makes it difficult to interpret the results. Answers 1-5 belong to the category 'In Labor Force', with 0-3 being 'employed' (but 0 unspecified), 4 is also confusing & 5 the only official 'unemployed' label and answers 6-10 belong to the category 'Not in Labor Force' and 11-12, 97-99 are invalid answers. To simply things, I recoded the variable to 'Employed' (0-3) and 'Unemployed' (4-10). Conceptually, this is not at all ideal, I'd rather focus on people in the Labor Force and distinguish between the 'employed' and the involuntarily 'unemployed'. But I already have too many missing NA's in this CV. I cannot just drop the whole population of people that are not in the Labor Force. Especially the 'retired' account for a substantial part of the population.

```
# Missing values
#sapply(df, function(x) sum(is.na(x))) # Show missing values
df$Vote_For_Incu[df$Vote_For_Incu %in% 9999996:9999999] <- NA # recode as NA
df$Eco_Eval[df$Eco_Eval %in% 7:9] <- NA
df$Employment[df$Employment %in% c(11:12, 97:99)] <- NA
df$Partisan[df$Partisan %in% c(7:9)] <- NA
df$Sat_with_Dem[df$Sat_with_Dem %in% c(7:9)] <- NA
sapply(df, function(x) sum(is.na(x)))
```

```
     Eco_Eval Vote_For_Incu    Employment      Partisan  Sat_with_Dem
       157935        125395        124257         30123         23275
```

```
df_cleaned <- df %>% filter(!is.na(Vote_For_Incu)) # remove NA's in DV
#str(df_cleaned)
df_recode <- df_cleaned %>% mutate(
        Employment = ifelse(Employment %in% 0:3, 1, 0),
        Sat_with_Dem = ifelse(Sat_with_Dem == 6, 3, Sat_with_Dem))
table(df_recode$Employment)
```

```
     0      1
168341 102061
```

```
table(df_recode$Sat_with_Dem)
```

```
     1      2      3      4      5
 31041 129380   3030  70035  25685
```

```
# Impute missing values for Partisan and Sat_with_Dem (few NA's)
df_recode$Partisan[is.na(df_recode$Partisan)] <-
  median(df_recode$Partisan, na.rm = TRUE)  # Median imputation
df_recode$Sat_with_Dem[is.na(df_recode$Sat_with_Dem)] <-
  median(df_recode$Sat_with_Dem, na.rm = TRUE)
# Impute missing values for Employment and Eco_Eval (many NA's) using pmm
```

```r
df_impute <- mice(df_recode, method = c(
  "pmm",        # Eco_Eval: Predictive Mean Matching
  "",           # Vote_For_Incu: Already cleaned
  "pmm",        # Employment
  "",           # Partisan: Already imputed
  ""            # Sat_with_Dem
), m = 5, seed = 123)
```

```
 iter imp variable
  1   1  Eco_Eval
  1   2  Eco_Eval
  1   3  Eco_Eval
  1   4  Eco_Eval
  1   5  Eco_Eval
  2   1  Eco_Eval
  2   2  Eco_Eval
  2   3  Eco_Eval
  2   4  Eco_Eval
  2   5  Eco_Eval
  3   1  Eco_Eval
  3   2  Eco_Eval
  3   3  Eco_Eval
  3   4  Eco_Eval
  3   5  Eco_Eval
  4   1  Eco_Eval
  4   2  Eco_Eval
  4   3  Eco_Eval
  4   4  Eco_Eval
  4   5  Eco_Eval
  5   1  Eco_Eval
  5   2  Eco_Eval
  5   3  Eco_Eval
  5   4  Eco_Eval
  5   5  Eco_Eval
```

```r
df_complete <- complete(df_impute, 1) # Extract completed dataset (1st imputed)
colSums(is.na(df_complete))
```

```
     Eco_Eval Vote_For_Incu    Employment      Partisan  Sat_with_Dem
            0             0             0             0             0
```

```
fit_temp <- lm(Vote_For_Incu ~ Eco_Eval + Employment + Partisan + Sat_with_Dem, data = df_co
vif(fit_temp)
```

```
    Eco_Eval   Employment     Partisan Sat_with_Dem 
    1.068807     1.001184     1.010269     1.066743 
```

All values are close to 1, indicating minimal or no multicollinearity among the predictors. The model is ready for interpretation.

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.

```
# OLS multiple regression model
Simple_model <- lm(Vote_For_Incu ~ Eco_Eval, data = df_complete)
#summary(Simple_model) # Only one predictor
OLS_model <- lm(Vote_For_Incu ~ Eco_Eval + Employment + Partisan + Sat_with_Dem, data = df_c
#summary(OLS_model) # Multiple predictors
stargazer(Simple_model, OLS_model, type = "text")
```

```
================================================================================
                                        Dependent variable:                     
                    ------------------------------------------------------------
                                           Vote_For_Incu                        
                                 (1)                            (2)             
--------------------------------------------------------------------------------
Eco_Eval                      -0.089***                      -0.080***          
                               (0.001)                        (0.001)           
                                                                                
Employment                                                    -0.002            
                                                               (0.002)           
                                                                                
Partisan                                                      0.023***          
                                                               (0.002)           
```

```
Sat_with_Dem                                                 -0.044***
                                                              (0.001)

Constant                       0.684***                      0.762***
                               (0.002)                        (0.003)

--------------------------------------------------------------------------------
Observations                   270,402                        270,402
R2                              0.082                          0.094
Adjusted R2                     0.082                          0.094
Residual Std. Error      0.470 (df = 270400)            0.466 (df = 270397)
F Statistic         24,051.740*** (df = 1; 270400) 7,008.255*** (df = 4; 270397)
================================================================================
Note:                                                 *p<0.1; **p<0.05; ***p<0.01
```

Important Note!!! I did not inverse the scale order of the IV and the CV 'Sat_with_Dem', so their relationship with the DV is negative, not positive. This is not a problem, but it is important to be aware of it when interpreting the results. If a higher value on the Eco_Eval variable represents a more negative economic evaluation, then a negative coefficient here actually aligns with the initial hypothesis that more positive economic evaluations would lead to greater likelihood of voting for the incumbent. In other words, better evaluations of the economy (lower Eco_Eval values) are related to higher support for the incumbent.

Simple Model: The first model includes only the independent variable Eco_Eval. The coefficient for Eco_Eval is negative and highly significant ($p<0.01$). This indicates that more pessimistic views on the state of the economy are associated with lower predicted values of Vote_For_Incu.

Full Model: In this model, the DV (Vote_For_Incu) is a binary indicator of whether the individual voted for the incumbent. ***The coefficients can be interpreted as changes in the probability of voting for the incumbent associated with a one-unit change in the explanatory variables, holding other variables constant.*** Economic Evaluation (-0.08): This suggests that a one-unit increase in the respondent's economic evaluation measure (more negative sentiment) is associated with about a 8 percentage point decrease in the probability of voting for the incumbent. The strong negative and highly statistically significant coefficient ($p< 0.001$) indicates that as negative economic evaluations decrease, the likelihood of supporting the incumbent increases.

Employment (-0.0023): This coefficient is not statistically significant ($p > 0.05$), meaning changes in the Employment variable do not reliably predict changes in the probability of voting for the incumbent. Effectively, employment, as measured here, doesn't have a clear linear relationship with incumbent support in this sample. However, it should be noted that the variable has some inherent flaws, such as the large amount of missing values and the conceptually flawed operationalization of employment status, especially as it relates to the decision of including people who are not part of the labor force. The hypotheses about unemployment (and with that maybe being pessimistic and resentful) having a negative effect on the personal evaluation of the incumbent's performance and the perceived state of the economy, does not really apply to the majority of people who are not part of the labor force for all sorts of reasons. It can be assumed that a lot of them are not involuntarily unemployed (or specifically: not in the labor force) and are therefore less likely to be disenchanted about their social and economic status. Thus, their status should as such not affect the IV and the DV as much.

Partisan (0.0223): This indicates that a one-unit increase in the partisan measure is associated with a 2.23 percentage point increase in the probability of voting for the incumbent. The positive and statistically significant effect ($p < 0.001$) suggests that stronger partisanship (maybe favoring the incumbent's party) boosts the likelihood of supporting that incumbent. As previously mentioned, conceptually, the variable is not ideal, as it does not exactly represent the theory of partisanship and how it connects to the perception of the incumbent's performance and the perceived state of the economy as it relates to the respondent's and the incumbent's party colors. However, I'm still glad I included it, as it turns out to be a good control variable for the model and even in its simpler form, there might be something to it from a theoretical standpoint.

Sat_with_Dem (-0.044): A one-unit increase in Dis!-satisfaction with democracy is associated with a 4.4 percentage point decrease in the probability of voting for the incumbent. This effect is strongly statistically significant ($p < 0.001$). Again, one needs to keep in mind that the scale of this variable is inverted, so higher values indicate lower satisfaction with democracy. The negative coefficient suggests that higher satisfaction with democracy is associated with a greater likelihood of supporting the incumbent. This is in line with the expectation that dissatisfaction with the political system could lead to a lower evaluation of the incumbent's performance and the general state of the economy.

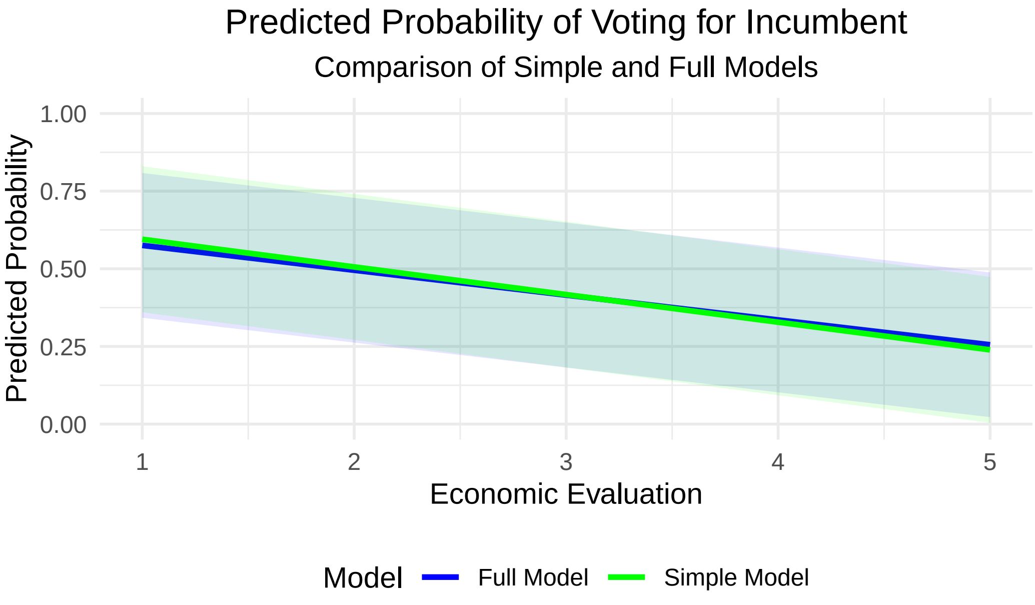
Both models show highly significant F-Statistics, and the R-squared increases from about 82 to 94% with the inclusion of the additional predictors, indicating a modest improvement in explanatory power when controlling for employment status, partisanship, and satisfaction with democracy.

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

```
# creating sequence of economic evaluation values
eco_eval_seq <- seq(min(df_complete$Eco_Eval, na.rm = TRUE),
                    max(df_complete$Eco_Eval, na.rm = TRUE),
                    length.out = 100)
predicted_probs <- data.frame( # calculating predicted prob for both models
  Eco_Eval = eco_eval_seq,
  # Full model (Model 2)
  full_model = 0.762 + # intercept
```

```r
    (-0.080 * eco_eval_seq) + # Eco_Eval coefficient
    (-0.002 * mean(df_complete$Employment)) + # Employment at mean
    (0.023 * mean(df_complete$Partisan)) + # Partisan at mean
    (-0.044 * mean(df_complete$Sat_with_Dem)), # Sat_with_Dem at mean
  # Simple model (Model 1)
  simple_model = 0.684 + # intercept
    (-0.089 * eco_eval_seq)) # Eco_Eval coefficient
ggplot(predicted_probs) + # Creating the plot
  geom_line(aes(x = Eco_Eval, y = full_model, color = "Full Model"), size = 1) +
  geom_ribbon(aes(x = Eco_Eval,   # full model
                  ymin = full_model - 0.466/2,
                  ymax = full_model + 0.466/2),
              alpha = 0.1, fill = "blue") +
  geom_line(aes(x = Eco_Eval, y = simple_model, color = "Simple Model"), size = 1) +
  geom_ribbon(aes(x = Eco_Eval, # simple model
                  ymin = simple_model - 0.470/2,
                  ymax = simple_model + 0.470/2),
              alpha = 0.1, fill = "green") +
  scale_color_manual(values = c("Full Model" = "blue", "Simple Model" = "green")) +
  labs(title = "Predicted Probability of Voting for Incumbent",
       subtitle = "Comparison of Simple and Full Models",
       x = "Economic Evaluation",
       y = "Predicted Probability",
       color = "Model",
       caption = "Shaded areas represent 95% confidence intervals") +
  theme_minimal() +
  scale_y_continuous(limits = c(0, 1)) +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "bottom")
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Shaded areas represent 95% confidence intervals

The visualization supports the regression results by clearly depicting the negative association between (negative!) economic evaluations and incumbent support. The downward-sloping lines indicate that as the economic evaluations become less favorable (higher values on the x-axis), the probability of voting for the incumbent decreases. This matches the regression coefficient for Eco_Eval in both models. The blue line represents the full model, which includes the additional predictors, while the green line represents the simple model with only the economic evaluation variable. The shaded areas around the lines represent the 95% confidence intervals for the predicted probabilities. They show that the model's predictions are more precise for intermediate values of economic evaluation but slightly less certain at the extremes.

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 CSESIMD 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!