Replicating a Experimental Study (THE I)

Research Design and Methods in Quantitative Research - Fall 2024

Álvaro Canalejo-Molero

2024-10-31

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**. 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: Read the Assigned Paper and Download the Replication Files

You will need to download and read the paper *Instrumentally Inclusive: The Political Psychology of Homonationalism*.

When you have read the paper, look in their replication files for the necessary files to replicate study 1. In particular, locate and download:

- Data file: `study1_data.csv`
- R Script: `study1.R`

The replication files provide no codebook for the data, so you will need to use the R script to navigate it and locate the relevant variables.

Exercises

Exercise 1: Summary of the Paper and Main Findings

Provide a brief summary of the paper you are replicating. Describe the main findings, especially those related to Study 1.

Exercise 2: Data Preparation and Exploration

Use the data file `study1_data.csv` to begin the replication process. Identify and describe the experimental variables (i.e., treatment and immigration attitudes) and provide visualizations of their distribution.

Then, select up to four covariates (e.g., gender, age, etc.) and plot their distribution too.

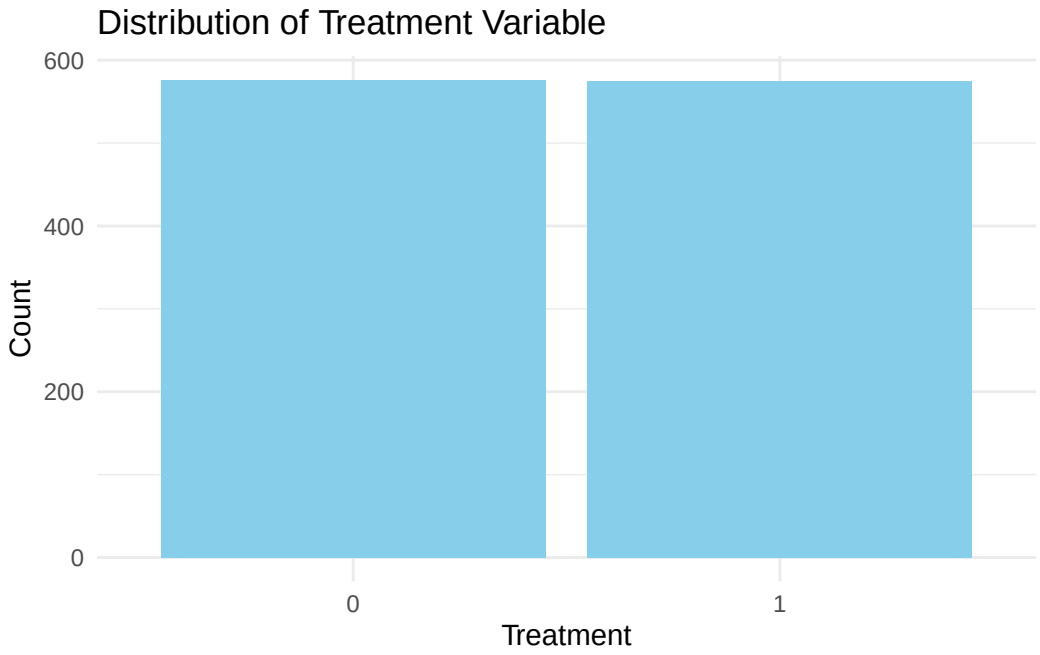
If necessary, clean or transform variables. Document any changes.

```
# Load necessary packages
library(tidyverse) # tidyverse environment
library(ggplot2) # nice plots

# Load the data
data <- read_csv("materials/study1_data.csv")

# Subset the data
data_subset <- data |>
  dplyr::select(
         support, # outcome variable
```

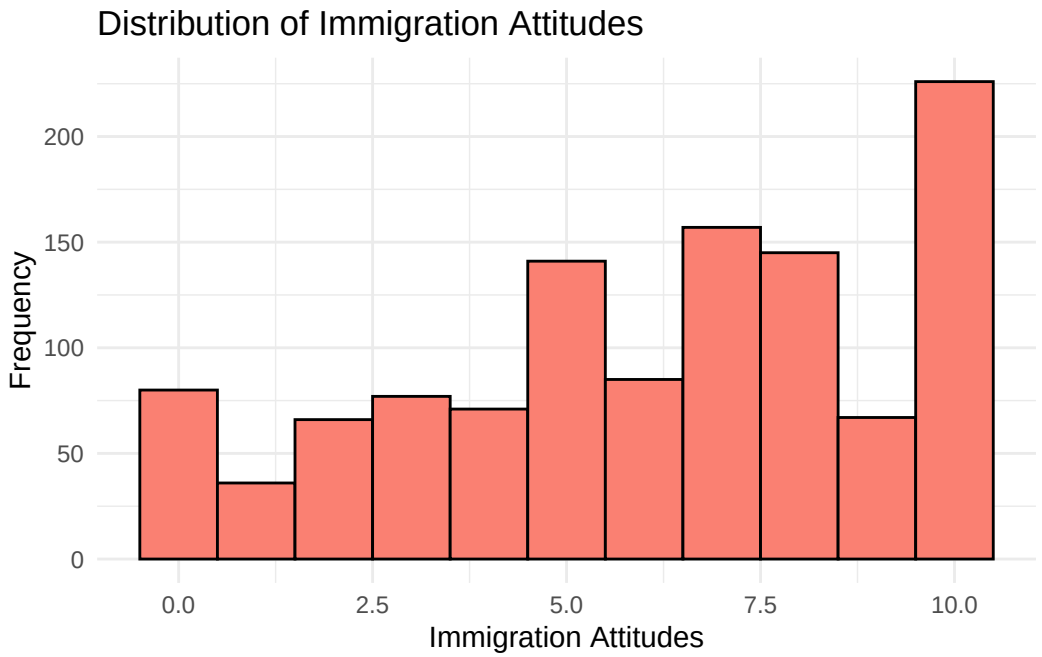
```
         treatment, # treatment variable
         imm_1, # conditional variable
         age,
         gender,
         degree,
         nonwhite
         ) |>
  # Factorizing variables
  mutate(treatment_fct = as.factor(treatment),
         gender_fct = as.factor(gender),
         gender = as.numeric(gender_fct),
         degree_fct = as.factor(degree),
         nonwhite_fct = as.factor(nonwhite),
         nonwhite = as.numeric(nonwhite_fct))

# Display summary of main variables
summary(data_subset)
```

```
    support         treatment          imm_1              age
 Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   :18.00
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 4.000   1st Qu.:35.00
 Median :1.0000   Median :0.0000   Median : 7.000   Median :48.00
 Mean   :0.6533   Mean   :0.4996   Mean   : 6.099   Mean   :47.52
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 9.000   3rd Qu.:62.00
 Max.   :1.0000   Max.   :1.0000   Max.   :10.000   Max.   :88.00
                                                    NA's   :3
     gender          degree          nonwhite     treatment_fct gender_fct
 Min.   :1.000   Min.   :0.0000   Min.   :1.000   0:576         Man  :549
 1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000   1:575         Woman:599
 Median :2.000   Median :0.0000   Median :1.000                 NA's :  3
 Mean   :1.522   Mean   :0.4294   Mean   :1.169
 3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:1.000
 Max.   :2.000   Max.   :1.0000   Max.   :2.000
 NA's   :3       NA's   :3
 degree_fct nonwhite_fct
 0   :655   0:956
 1   :493   1:195
 NA's:  3
```

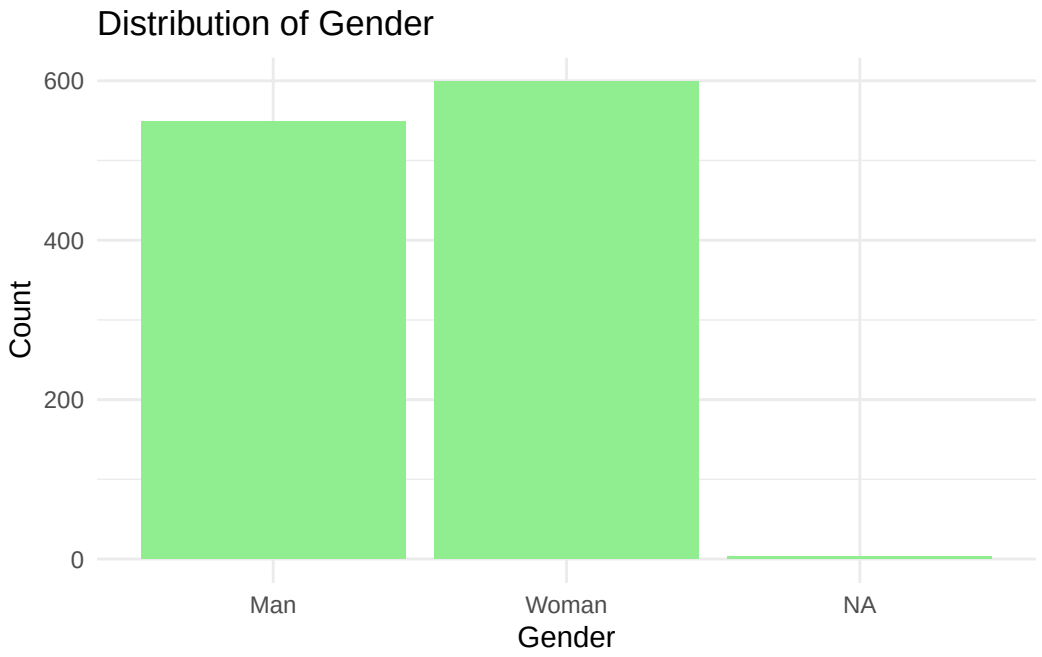
```r
# Visualize the distribution of the treatment variable
ggplot(data_subset, aes(x = treatment_fct)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Distribution of Treatment Variable",
       x = "Treatment",
       y = "Count") +
  theme_minimal()
```



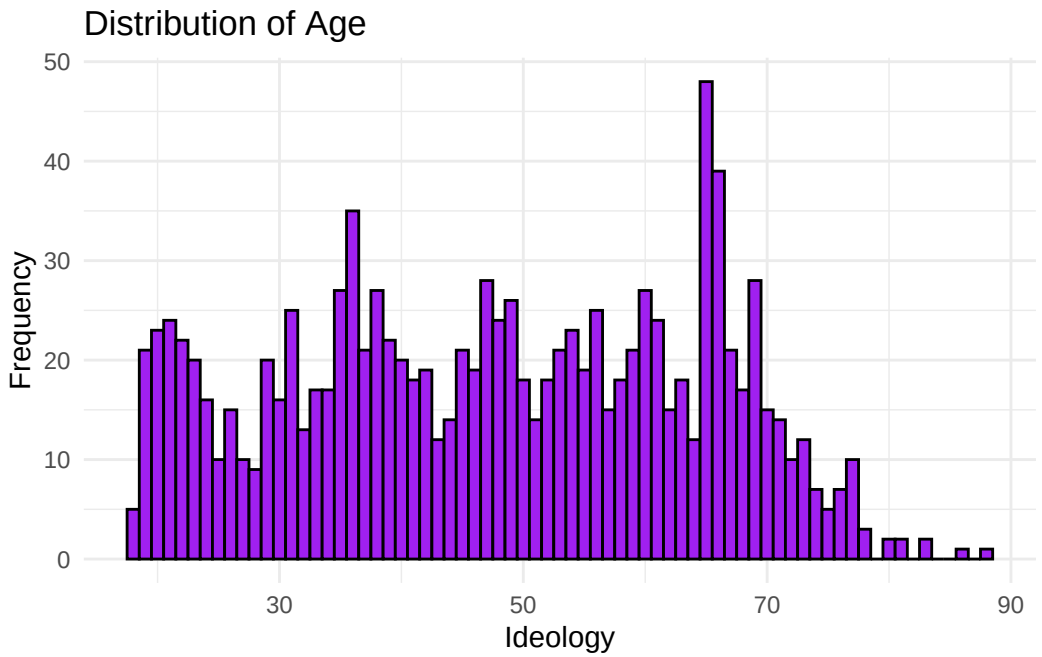

```r
# Visualize the distribution of the immigration attitudes variable
ggplot(data_subset, aes(x = imm_1)) +
  geom_histogram(binwidth = 1, fill = "salmon", color = "black") +
  labs(title = "Distribution of Immigration Attitudes",
       x = "Immigration Attitudes",
       y = "Frequency") +
  theme_minimal()
```



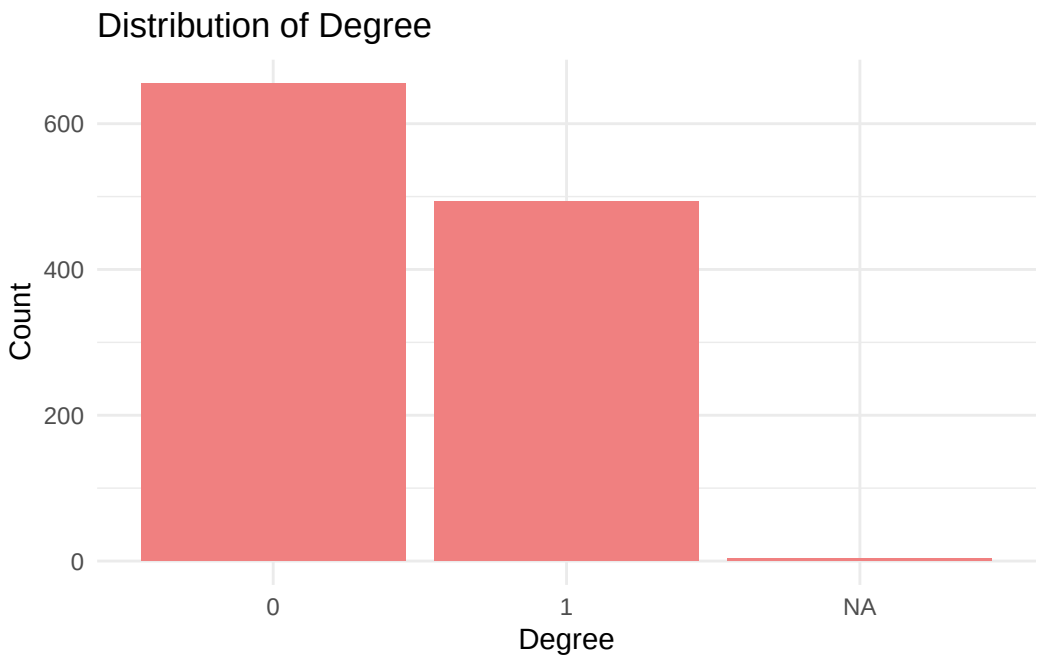

```
# Select and visualize distributions for up to four covariates

## Gender
ggplot(data_subset, aes(x = gender_fct)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Distribution of Gender",
       x = "Gender",
       y = "Count") +
  theme_minimal()
```



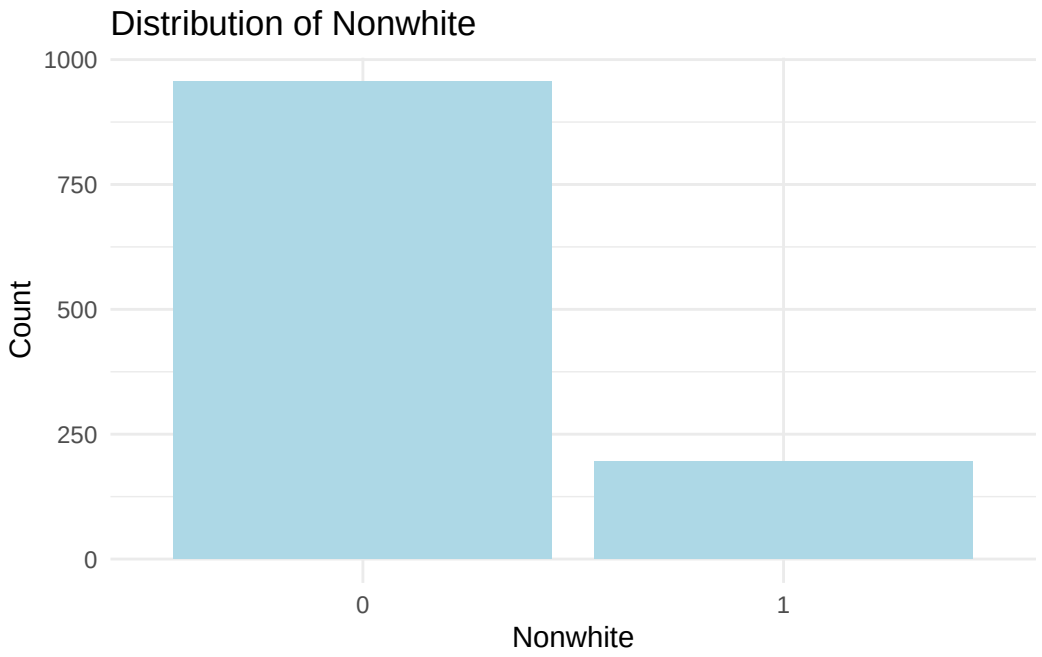

```
## Age
ggplot(data_subset, aes(x = age)) +
  geom_histogram(binwidth = 1, fill = "purple", color = "black") +
  labs(title = "Distribution of Age",
       x = "Ideology",
       y = "Frequency") +
  theme_minimal()
```




```
## Degree
ggplot(data_subset, aes(x = degree_fct)) +
  geom_bar(fill = "lightcoral") +
  labs(title = "Distribution of Degree",
       x = "Degree",
       y = "Count") +
  theme_minimal()
```




```
## Nonwhite
ggplot(data_subset, aes(x = nonwhite_fct)) +
  geom_bar(fill = "lightblue") +
  labs(title = "Distribution of Nonwhite",
       x = "Nonwhite",
       y = "Count") +
  theme_minimal()
```



Exercise 3: Covariate Balance

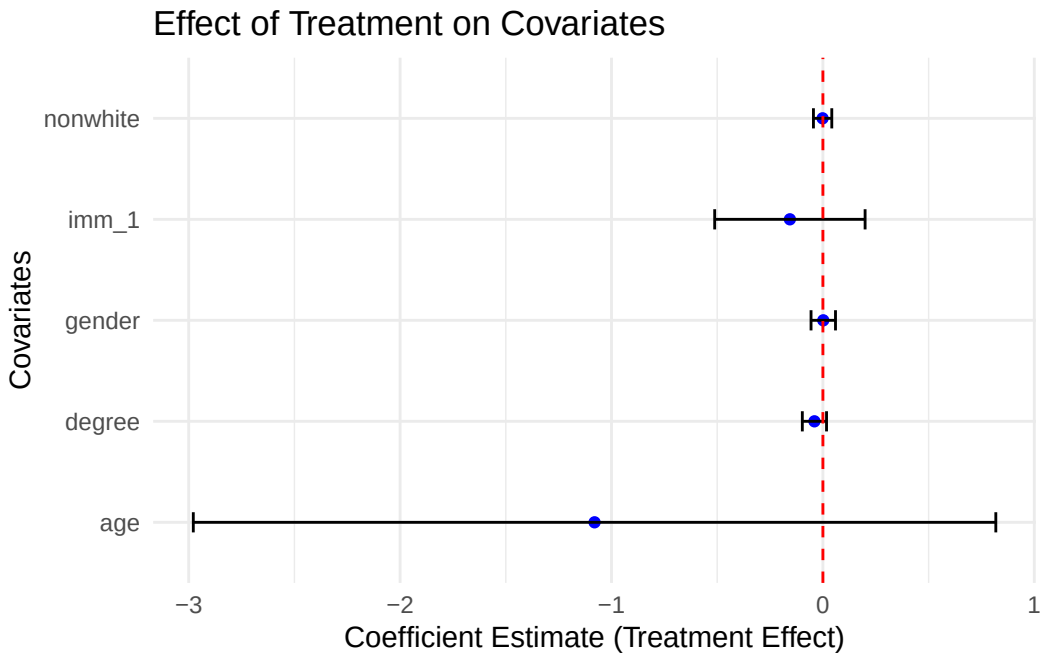
Check for balance across covariates and report the results in a table and a plot.

Explain your findings. Why would you expect randomization to lead to balance across covariates?

```
# Load necessary packages
library(tableone) # tables for covariate balance
library(broom) # tidy data from models
library(purrr) # + tidy data from models

# Define covariates to check for balance
covariates <- c("imm_1", "age", "gender", "degree", "nonwhite")

# Create a balance table by treatment
balance_table <- CreateTableOne(vars = covariates,
                                strata = "treatment_fct",
                                data = data_subset)

# Print the balance table
print(balance_table, showAllLevels = TRUE)
```

```
                     Stratified by treatment_fct
                      level 0             1              p      test
  n                           576           575
  imm_1 (mean (SD))          6.18 (3.06)   6.02 (3.10)   0.390
  age (mean (SD))           48.06 (15.95) 46.98 (16.85)  0.265
  gender (mean (SD))         1.52 (0.50)   1.52 (0.50)   0.953
  degree (mean (SD))         0.45 (0.50)   0.41 (0.49)   0.171
  nonwhite (mean (SD))       1.17 (0.38)   1.17 (0.37)   0.948
```

```r
# Run regressions and extract coefficients
balance_results_df <- map_dfr(covariates, function(covariate) {
  model <- lm(as.formula(paste(covariate, "~ treatment_fct")),
              data = data_subset)
  tidy(model) |>
    filter(term == "treatment_fct1") |>  # Modify if different factor levels
    mutate(covariate = covariate)
})

# Plot the coefficients for treatment effects on covariates
ggplot(balance_results_df, aes(x = estimate, y = covariate)) +
  geom_point(color = "blue") +
  geom_errorbarh(aes(xmin = estimate - 1.96 * std.error,
                     xmax = estimate + 1.96 * std.error), height = 0.2) +
  labs(title = "Effect of Treatment on Covariates",
       x = "Coefficient Estimate (Treatment Effect)",
       y = "Covariates") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal()
```



Exercise 4 (additional): Estimate Treatment Effect

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

Estimate the average effect of the treatment on the outcome variable `support` conditional on the pre-treatment immigration attitudes `imm_1`. For this, use an interaction term in an OLS regression model. Compare your results to those in the original paper.

Then, repeat this analysis with three iteratively smaller random samples of the treatment (n = 200, n = 100, n = 10) and control groups (n = 200, n = 100, n = 10); total N = 400, 200, and 20, respectively. Explain your findings.

Finally, discuss how sample size impacts the results and what this implies about the role of randomization for selection bias.

```
# Set seed for replication
set.seed(123)

# Run an OLS model with int. between treatment and immigration attitudes
model_full <- lm(support ~ treatment_fct * imm_1, data = data_subset)
summary(model_full)
```

```
Call:
lm(formula = support ~ treatment_fct * imm_1, data = data_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8926 -0.4344  0.1519  0.3038  0.7620

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.238052   0.041717   5.706 1.47e-08 ***
treatment_fct1        0.148216   0.058093   2.551   0.0109 *
imm_1                 0.065453   0.006054  10.811  < 2e-16 ***
treatment_fct1:imm_1 -0.019267   0.008503  -2.266   0.0236 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4437 on 1147 degrees of freedom
Multiple R-squared:  0.134, Adjusted R-squared:  0.1317
F-statistic: 59.14 on 3 and 1147 DF,  p-value: < 2.2e-16
```

```r
# Set sample sizes for treatment and control groups
sample_sizes <- c(200, 100, 10)

# Function to sample and fit model
sample_and_fit <- function(n) {
  sample_data <- data_subset |>
    group_by(treatment_fct) |>
    sample_n(n) |>
    ungroup()

  # Fit the model on the sample data
  model_sample <- lm(support ~ treatment_fct * imm_1, data = sample_data)

  # Summarize the model and return coefficients
  tidy(model_sample) |>
    filter(term == "treatment_fct1:imm_1") |> # Interaction term for CATE
    mutate(sample_size = n * 2) # Total sample size
}

# Apply function for each sample size and combine results
results_samples <- map_dfr(sample_sizes, sample_and_fit)
print(results_samples)
```

```
# A tibble: 3 x 6
  term                 estimate std.error statistic p.value sample_size
  <chr>                   <dbl>     <dbl>     <dbl>   <dbl>       <dbl>
1 treatment_fct1:imm_1  -0.0217    0.0146    -1.48    0.139         400
2 treatment_fct1:imm_1  -0.0286    0.0199    -1.44    0.152         200
3 treatment_fct1:imm_1  -0.0212    0.0890    -0.238   0.815          20
```

Submission guidelines

Please submit both the PDF file and the `.qmd` file. Both files should report all the code used for analysis and annotations explaining each step.

The name of the files must follow the structure *take-home_exercise_i_YOURSURNAME(S).pdf* and *take-home_exercise_i_YOURSURNAME(S).qmd*, respectively. They should be upload to the folder *Students responses/Take-home exercises/Take-home exercise I* in OLAT.

Deadline: **14.11.24**

References

Turnbull-Dugarte, S. J., & Ortega, A. L. (2024). Instrumentally inclusive: the political psychology of homonationalism. *American Political Science Review, 118(3)*, 1360-1378.