Replicating an Observational Study (THE II)

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 skipp 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: Read the Assigned Paper and Download the Replication Files

You will need to read the paper *Wealth of Tongues: Why Peripheral Regions Vote for the Radical Right in Germany*, which was assigned as a supplementary reading for ***session 09***.

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

- Data file: `data_main.rds`
- R Script: `tab_1_A4_A6_A7_A11.R`

The replication files provide no codebook for the data (as you see, this is often the case). Therefore, you will need to use the R script to navigate the dataset and locate the relevant variables.

Exercise

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 the effect of dialectal distance on voting for AfD at the county-level.

Exercise 2: Data Preparation and Exploration

Use the data file `data_main.rds` to begin the replication process. The main independent variable is dialectal distance to Hannover (`hannover_dist`). The dependent variables is county-level AfD vote share (`afd_party_17`). Please, provide a visualization of their distributions.

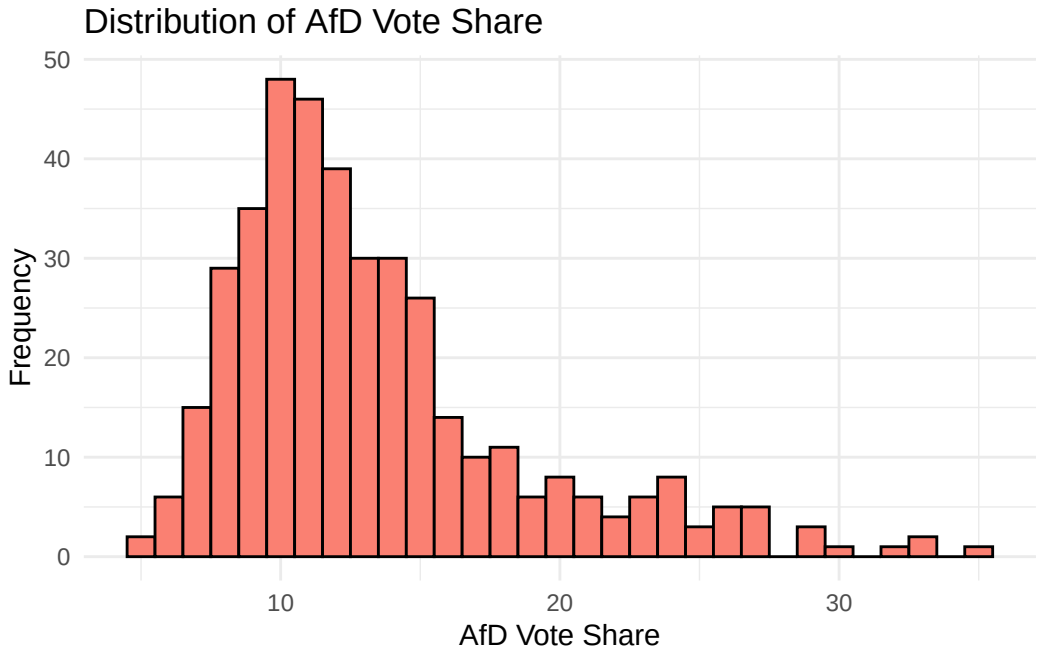
Then, generate a scatter plot of the relationship between dialectal distance and AfD vote share.

Please comment your findings.

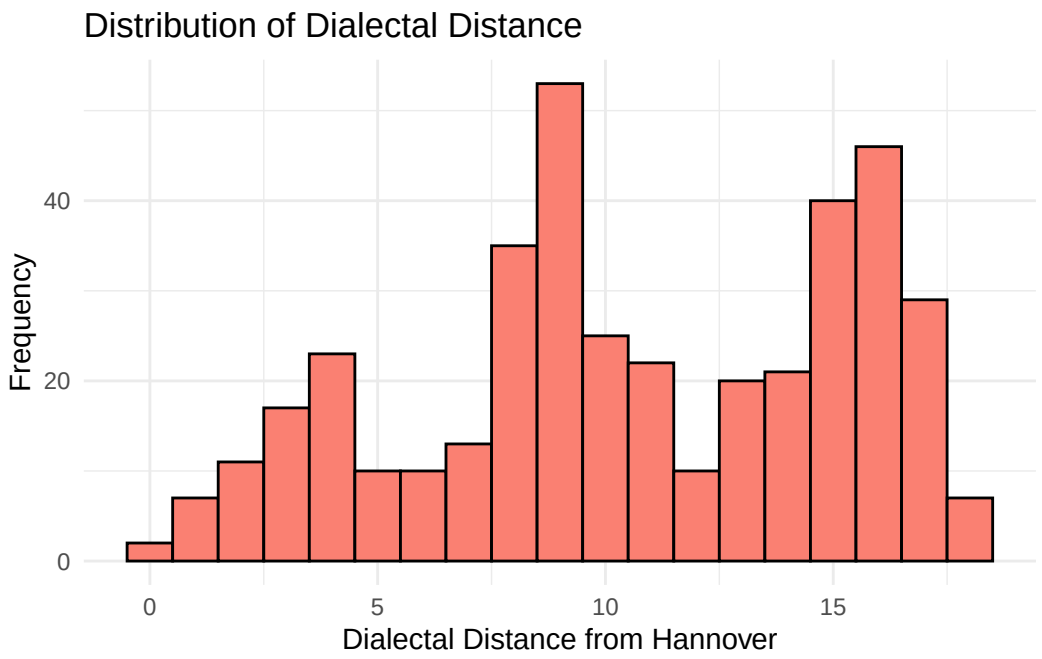
```
# Load necessary packages
library(tidyverse) # tidyverse environment
library(ggplot2) # nice plots

# Load the data
data <- read_rds('materials/data_main.rds')
```

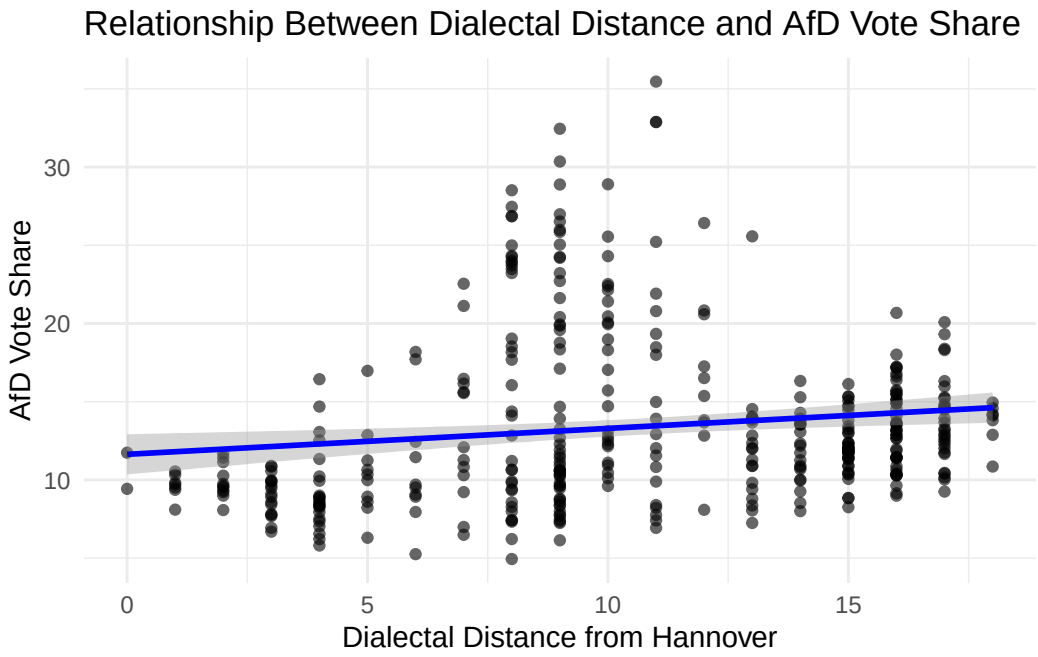
```r
# Visualize the distribution of AfD vote share
ggplot(data, aes(x = afd_party_17)) +
  geom_histogram(binwidth = 1, fill = "salmon", color = "black") +
  labs(title = "Distribution of AfD Vote Share",
       x = "AfD Vote Share",
       y = "Frequency") +
  theme_minimal()
```




```r
# Visualize the distribution of dialectal distance
ggplot(data, aes(x = hannover_dist)) +
  geom_histogram(binwidth = 1, fill = "salmon", color = "black") +
  labs(title = "Distribution of Dialectal Distance",
       x = "Dialectal Distance from Hannover",
       y = "Frequency") +
  theme_minimal()
```




```
# Scatter plot of main relationship
ggplot(data, aes(x = hannover_dist, y = afd_party_17)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", color = "blue", se = TRUE) +
  labs(title = "Relationship Between Dialectal Distance and AfD Vote Share",
       x = "Dialectal Distance from Hannover",
       y = "AfD Vote Share") +
  theme_minimal()
```



Exercise 3: Replicate Main Specification

Now, use the same code as the authors in `tab_1_A4_A6_A7_A11.R` to replicate `m0`. This is their baseline model without state fixed effects (`state`), in which they just cluster the standard errors by county (`ags_2017`). Comment whether the replication was successful.

Then, replicate the models adding state fixed effects without any additional covariate and comment the results. How are they different from your previous specification? And from the authors' reported results? Please explain why you think the results differ.

PS: Please remember to always scale the main independent variable. You can use the function `scale()`, as the authors do in their code. For fixed-effects models, follow the authors and use the function `felm` from the package `lfe`.

```
# Load necessary packages
library(lfe) # to run fixed effects models

# Replication authors' main specification
m0 <- felm(afd_party_17 ~ scale(hannover_dist) | 0 |0 | ags_2017,
           data = data)
summary(m0)
```

```
Call:
   felm(formula = afd_party_17 ~ scale(hannover_dist) | 0 | 0 |      ags_2017, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-8.021 -3.416 -1.647  1.336 22.006

Coefficients:
                     Estimate Cluster s.e. t value Pr(>|t|)
(Intercept)           13.3928       0.2641  50.716  < 2e-16 ***
scale(hannover_dist)   0.7779       0.1700   4.576 6.34e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.282 on 398 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared(full model): 0.02133   Adjusted R-squared: 0.01887
Multiple R-squared(proj model): 0.02133   Adjusted R-squared: 0.01887
F-statistic(full model, *iid*):8.676 on 1 and 398 DF, p-value: 0.003414
F-statistic(proj model): 20.94 on 1 and 398 DF, p-value: 6.335e-06
```

```r
# Replication with fixed-effects
m0_fe <- felm(afd_party_17 ~ scale(hannover_dist) | state |0 | ags_2017,
           data = data)
summary(m0_fe)
```

```
Call:
   felm(formula = afd_party_17 ~ scale(hannover_dist) | state |      0 | ags_2017, data = dat

Residuals:
    Min      1Q  Median      3Q     Max
-9.7660 -1.6666 -0.0135  1.6039  8.7812

Coefficients:
                     Estimate Cluster s.e. t value Pr(>|t|)
scale(hannover_dist)   1.5345       0.3495    4.39 1.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.573 on 384 degrees of freedom
  (1 observation deleted due to missingness)
```

```
Multiple R-squared(full model): 0.7759   Adjusted R-squared: 0.7672
Multiple R-squared(proj model): 0.04176   Adjusted R-squared: 0.004332
F-statistic(full model, *iid*):88.64 on 15 and 384 DF, p-value: < 2.2e-16
F-statistic(proj model): 19.27 on 1 and 384 DF, p-value: 1.466e-05
```

Exercise 4 (additional): Additional Replication

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

Now add **two** covariates to your baseline specification (i.e., without state-fixed effects). You can select any covariates from the authors' model `m01`. Justify your selection of covariates and then explain how the results differ from your previous specification.

Based on the results, do you think you selected useful covariates to control for?

Finally, add state fixed-effects to your two-control variables specification. Compare your results across specifications, including those reported by the authors. What conclusions do you draw about the role of control variables to isolate causal effects?

```r
# Replication baseline specification (without FEs) with two covariates
m0_controls <- felm(afd_party_17 ~ scale(hannover_dist) + unemp_rate_tot +
             dist_to_state_capital | 0 |0 | ags_2017,
           data = data) # control variables can be different
summary(m0_controls)
```

```
Call:
   felm(formula = afd_party_17 ~ scale(hannover_dist) + unemp_rate_tot +      dist_to_state_

Residuals:
    Min      1Q  Median      3Q     Max
-8.7840 -2.6553 -0.3705  1.7489 18.3198

Coefficients:
                       Estimate Cluster s.e. t value Pr(>|t|)
(Intercept)            7.102089     0.618057  11.491   <2e-16 ***
scale(hannover_dist)   2.387349     0.186563  12.797   <2e-16 ***
unemp_rate_tot         1.023672     0.093605  10.936   <2e-16 ***
dist_to_state_capital -0.005027     0.003134  -1.604    0.109
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.398 on 388 degrees of freedom
```

```
  (9 observations deleted due to missingness)
Multiple R-squared(full model): 0.321   Adjusted R-squared: 0.3158
Multiple R-squared(proj model): 0.321   Adjusted R-squared: 0.3158
F-statistic(full model, *iid*):61.15 on 3 and 388 DF, p-value: < 2.2e-16
F-statistic(proj model): 62.16 on 3 and 388 DF, p-value: < 2.2e-16
```

```r
# Replication baseline specification with FEs and the same two covariates
m0_fe_controls <- felm(afd_party_17 ~ scale(hannover_dist) + unemp_rate_tot +
            dist_to_state_capital | state |0 | ags_2017,
          data = data) # control variables can be different
summary(m0_fe_controls)
```

```
Call:
   felm(formula = afd_party_17 ~ scale(hannover_dist) + unemp_rate_tot +      dist_to_state_

Residuals:
     Min       1Q   Median       3Q      Max
-10.3980  -1.6065   0.0521   1.4819   9.2258

Coefficients:
                       Estimate Cluster s.e. t value Pr(>|t|)
scale(hannover_dist)   1.455966     0.363368   4.007 7.42e-05 ***
unemp_rate_tot         0.341444     0.068539   4.982 9.64e-07 ***
dist_to_state_capital -0.004483     0.002554  -1.755     0.08 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.489 on 375 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared(full model): 0.7899   Adjusted R-squared: 0.7809
Multiple R-squared(proj model): 0.1079   Adjusted R-squared: 0.0698
F-statistic(full model, *iid*): 88.1 on 16 and 375 DF, p-value: < 2.2e-16
F-statistic(proj model): 14.22 on 3 and 375 DF, p-value: 8.506e-09
```

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_ii_YOURSURNAME(S).pdf* and *take-home_exercise_ii_YOURSURNAME(S).qmd*, respectively. They should be upload to the folder *Students responses/Take-home exercises/Take-home exercise II* in OLAT.

Deadline: **28.11.24**

References

Ziblatt, D., Hilbig, H., & Bischof, D. (2024). Wealth of tongues: Why peripheral regions vote for the radical right in germany. American Political Science Review, 118(3), 1480-1496.