The Effects of COVID-19 Infection on Opposition to COVID-19 Policies: Evidence from the U.S. Congress

Replication Instructions

Zach Dickson (LSE) and T. Murat Yildirim (Stavanger)

2024-04-21

Replication files for: *The Effects of COVID-19 Infection on Opposition to COVID-19 Policies: Evidence from the U.S. Congress*

Authors: Zach Dickson (LSE) and T. Murat Yildirim (Stavanger)

Journal: *Political Communication*

To run the entire analysis, you will need to have R and Python installed on your machine.

There are eight files necessary for replication that are in the main directory:

1. `master.py` - This file runs the entire analysis.
2. `analysis.R` - This file runs all R code
3. `figures.py` - This file runs all Python code
4. `analysis_data.csv` - The primary dataset used in the analysis
5. `covid_infections.csv` - The data file for COVID infections in legislators
6. `validation_tweets.csv` - The data file for the held-out tweets used to validate the language model
7. `py_requirements.txt` - Python library requirements
8. `requirements.R` - R library requirements

Additionally, there are two folders that contain the individuals scripts for the analysis (`individual_files`) and all the compiled files (`compiled_files`). If these are not of interest, you can delete both folders and just compile the entire replication using the `master.py` file. Clone the repo, navigate to the new directory and run the following in your terminal:

```
python3 master.py
```

Time to Run: 49 minutes on 20 core, 64GB RAM machine.

Notes:

- The master file calls the two analysis files in R and Python. The R file estimates the primary model using matrix completion, and the Python file creates the figures and tables. These files will save the results in the main directory.
- There are also two requirements files for each language.
- There are two folders in which the individual analysis files are broken up by the type of analysis (`individual_files`) and one that contains all compiled files (`compiled_files`). These files are further detailed below.
- In several cases I estimate the models in R and then save the results as a csv file. I then use the csv file to create the figures in Python using the `figures.ipynb` file.
- I tried to set seeds where possible to ensure reproducibility, but many of the models are stochastic and may not be exactly the same as in the paper.

Data

All data are in .csv format. The data includes the following files:

- `analysis_data.csv` contains the primary dataset used in the analysis
- `covid_infections.csv` contains the data file used to create figure 1 in the paper (COVID infections in legislators)
- `figure3.csv` contains the data file used to create figure 3 in the paper (effects of COVID infection on opposition 4 weeks before and after the infection)
- `figure4.csv` contains the data file used to create figure 4 in the paper (exit effects of COVID infection on opposition)
- `figureA3.csv` contains the data file used to create figure A3 in the Appendix (effects of COVID infection on opposition 4 weeks before and after the infection using Congressional Press Releases)
- `validation_tweets.csv` contains the 1000 held-out tweets used to validate the language model with the model's predictions.

Code

- `analysis_MC.R` contains the code to reproduce the primary analysis in the paper using Matrix Completion methods.
- `figures.ipynb` contains the code to reproduce many of the figures and the descriptive statistics presented in the paper (Python notebook).
- `causal_forest.R` contains the code to reproduce estimation of heterogeneous treatment effects using the causal forest method.
- `robustness_check1.R` contains the code to reproduce the robustness check using matrix completion with infected legislators only.

- `robustness_check2.R` contains the code to reproduce the estimation of infection on the number of total tweets from legislators using matrix completion.
- `robustness_check3.R` contains the code to reproduce the estimation of effects of infection on opposition using Congressional Press releases using matrix completion.
- `robustness_check4.R` contains the code to reproduce the estimation of effects of infection on opposition using interactive fixed effects estimator.
- `robustness_check5.R` contains the code to reproduce the estimation of effects of infection on opposition using interactive fixed effects estimator with infected legislators only.
- `robustness_check6.R` contains the code to reproduce the estimation of effects of infection on total tweets using interactive fixed effects estimator.
- `robustness_check7.R` contains the code to reproduce the estimation of effects of infection on total tweets using interactive fixed effects estimator with Congressional Press Releases.

Tables

- `table1.tex` contains the table 1 in the paper (cumulative effects of COVID infection on opposition)
- `tableA1.tex` contains the table A1 in the Appendix (descriptive statistics for tweets)
- `tableA2.tex` contains the table A2 in the Appendix (CATE estimates using causal forest)
- `tableA3.tex` contains the table A3 in the Appendix (Estimation with infected legislators only - Robustness check)
- `tableA4.tex` contains the table A4 in the Appendix (Estimation of infections on the number of total tweets from legislators)
- `tableA5.tex` contains the table A5 in the Appendix (Estimation of effects of infection on opposition using Congressional Press releases)
- `tableA6.tex` contains the table A6 in the Appendix (Descriptive statistics for Congressional Press Releases)
- `tableA7.tex` contains the table A7 in the Appendix (Estimation of effects of infection on opposition using interactive fixed effects estimator)
- `tableA8.tex` contains the table A8 in the Appendix (Estimation of effects of infection on opposition using interactive fixed effects estimator with infected legislators only)
- `tableA9.tex` contains the table A9 in the Appendix (Estimation of effects of infection on total tweets using interactive fixed effects estimator)
- `tableA10.tex` contains the table A10 in the Appendix (Estimation of effects of infection on total tweets using interactive fixed effects estimator with Congressional Press Releases)

Figures

- `figure1.png` contains the figure 1 in the paper (COVID infections in legislators)
- `figure2.png` contains the figure 2 in the paper (opposition to COVID measures by legislators)
- `figure3.png` contains the figure 3 in the paper (effects of COVID infection on opposition 4 weeks before and after the infection)
- `figure4.png` contains the figure 4 in the paper (exit effects of COVID infection on opposition)
- `figureA1.png` contains the figure A1 in the Appendix (F1 scores for validation of language model)
- `figureA2.png` contains the figure A2 in the Appendix (Pre-trends equivalence tests for matrix completion method)
- `figureA3.png` contains the figure A3 in the Appendix (effects of COVID infection on opposition 4 weeks before and after the infection using Congressional Press Releases - Robustness check)

Additional Notes:

The fine-tuned BERT model used in the analysis is available at this link. The model can be loaded using the following code in Python:

```
# Load the necessary libraries
from transformers import BertTokenizer, BertForSequenceClassification, pipeline

# Load the model and tokenizer
model_name = 'z-dickson/US_politicians_covid_skepticism'
tokenizer = BertTokenizer.from_pretrained(model_name)
model = BertForSequenceClassification.from_pretrained(model_name)

# Load the model into a pipeline
classifier = pipeline('sentiment-analysis',
                      model=model,
                      tokenizer=tokenizer)

# Example usage
classifier("I am skeptical about COVID-19 measures")
```

Descriptions of the Data Sources

There are several sources of data that were used, which are detailed below:

1. The primary data source for legislators' tweets was the Twitter API. This has since been discontinued; however, the data can be collected using the `congresstweets` repo in github (https://github.com/alexlitel/congresstweets).

2. The secondary source of data for the outcome variable using Legislators' press releases came from the ProPublica Congress API (https://projects.propublica.org/api-docs/congress-api/statements/). An account is required to access the API.

3. The data on COVID-19 infections in legislators was collected from GovTrack (https://www.govtrack.us/covid-19). You can download that data from the website directly. Additionally, I provide the code in the `figures.ipynb` file to import the data directly from the website.

4. The data on COVID-19 infections in the general population (at the state level) was collected from the New York Times COVID-19 data repository (https://github.com/nytimes/covid-19-data). You can download the data from the website directly.