README

Paper: Causal Identification in Social Sciences using Observational Data

Author: Fırat Bilgel

The supplemental material for the article contains the Supplemental Online Appendix to the paper and the code and data to replicate all analyses.

The supplemental material consists of the following files:

id#	File name	Description
1	replication_LAWC_final.R	R script to replicate the simulation results of Tables 1 and 2 (Section 3.2)
2	replication_LAWC_impossible.do	do file to replicate the simulation results of Table 3 (Section 3.2)
3	replication_JRSD.R	R script to replicate the simulation results of Section 3.3
4	replication_MCAS.R	R script to replicate the simulation results of Table 4 (Section 3.4)
5	replication_SARR.R	R script to replicate the simulation results of Table 5 (Section 3.5)
6	replication_SARR.do	do file to replicate the IV probit results of Table 6 (Section 3.5)
7	northcarolina1.txt northcarolina2.txt	txt files of the North Carolina recidivism dataset
8	nc_recid (observational).csv	csv file of the observational distribution, $\Pr(Y \mid T)$ via Logistic regression using Causal Fusion software
9	nc_recid (experimental).csv	csv file of the experimental distribution, $\Pr(Y \mid do(T))$ via Logistic regression using Causal Fusion software
10	DAGs.tex	LaTeX codes to reproduce Figs. 2-8, 9.b, 10.b-c using tikz package
11	tradeoff.tex	LaTeX code to reproduce Fig. 1 using tikz package
12	DAGs_appendix.tex	LaTeX code to reproduce Figs. A.1, A.2 of the appendix using tikz package

All computations except for id #2 and #6 were performed using the statistical software R. This software is provided "as is" without warranty of any kind, expressed or implied. For questions and further error reports, contact bilgelf@mef.edu.tr. All computations for id #2 and #6 were performed using Stata 15 MP.

Instructions

1. Set the working directory at the top command line of each R script file in id #1,3, 4 and 5 to the working directory to which the uncompressed folder is placed.

2. Set the directories for the csv or dta files produced by the R script files in id #1,3, 4 and 5 to the working directory to which the uncompressed folder is placed.

R packages and versions

The script files in id #1, 3, 4 and 5 were tested using RStudio/2023.06.1+524 "Mountain Hydrangea" Release (547dcf86, 2023-07-06) for macOS Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko). The following packages are used:

Package	Version	Package	Version	Package	Version
haven	2.5.3	caret	6.0.94	modelsummary	1.4.3
mvtnorm	1.2.2	ggplot2	3.4.4	knitr	1.45
MASS	7.3.60	gridExtra	2.3	kableExtra	1.3.4
boot	1.3.28.1	endogeneity	2.1.3	lavaan	0.6.18
dplyr	1.1.4	tidyr	1.3.0	mediation	4.5.0
tidyverse	2.0.0	pbivnorm	0.6.0		

Variable descriptions

The R script file in id #1 generates the following variables in the simulation:

Variable	Description	Variable	Description
U1	family SES (multivalued)	Z2	quality of legal representation (5-level index)
U2	community wealth (multivalued)	X1	arrest circumstances (6-level index)
U3	legal system bias (5-level index)	X2	criminal history (years)
U4	judicial discretion (5-level index)	Treatment	access to legal aid (binary)
Z1	knowledge of the legal system (5-level index)	Y	wrongful conviction (binary)

The R script file in id #3 generates the following variables in the simulation:

Variable	Description	Variable	Description
U	judicial biases (5-level index)	J	judicial review panel (binary)
W	political pressure (5-level index)	M	mandatory sentencing guidelines (binary)
Z	judicial discretion (trichotomous)	Y	sentencing disparities (Theil index)
P	public perception of judicial fairness (trichotomous)		

The R script file in id #4 generates the following variables in the simulation:

Variable	Description
X	confounder (continuous)
D	media coverage (multivalued)
M	public opinion (trichotomous)
B	judge behavior (trichotomous)
Y	sentencing (months)

The R script file in id #5 uses the following variables:

Variable	Description	Variable	Description	Variable	Description
recid	recidivism (binary)	gender	gender (binary	married	marital status (binary)
prisontime	prison time (years)	alcohol	alcohol use (binary)	priors	prior criminal record (years)
race	race (binary)	drugs	hard drug use (binary)	felony	felony crime (binary)
age	age (years)	school	schooling (years)	property	property crime (binary)
				personal	personal crime (binary)

Files Organization

Figure/Table #	Program	Line number	Output file	Note
Table 1	replication_LAWC_final.R	1-56	-	Simulated dataset
	replication_LAWC_final.R	58-63, 77-114, 175-238		Bivariate probit ATT estimates, CI
Table 2	replication_LAWC_final.R	116-134, 145-172		Frontdoor ATT estimates, CI
Table 3	replication_LAWC_final.R	1-56	lawc_impossible.dta	input for the do file
	replication_LAWC_impossible.do	1-36	-	Custom table
Table 4	replication_MCAS.R	1-135	-	Custom table
Table 5	replication_SARR.R	1-50	nc_recid.csv	Custom table
Table 6	replication_SARR.do	1-13	ivprobit_SARR.tex	Custom table
Figure 1	tradeoff.tex	-	-	tikz package LaTeX codes
Figures 2-8, 9.b,10.b-c	DAGs.tex	-	-	tikz package LaTeX codes
Figures 9.a	-	-	pc_undirected.png	Custom drawing
Figure 10.a	-	-	rfci.png	Custom drawing
Figure 11	replication_SARR.R	59-169	nc_resid_logistic_BFPX_1.png	
Figures 1, 2	DAGs_appendix.tex	-	-	tikz package LaTeX codes