README FILE FOR SIMULATIONS IN PAPER

April 12, 2024

Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes

1 FILES

1. `NICS.Functions.R`: Incudes all functions needed to replicate the simulations in the paper. These include functions to generate data for the different models, compute the various estimators, and compute standard errors and confidence intervals.

2. `NICS.Main.R`: is the main file the do a run of simulations (typically, a row in each of the tables in the paper). This file calls the functions in `NICS.Functions.R`. The simulations where run via a script in a computer cluster and so the main values of the parameters were feeded from the script after reading their values from a series of .txt files listed below. This is visible in lines 11 to 25 in the file:

```
[11]   # Parametrics provided in the Script
[12]       args <- commandArgs(trailingOnly = TRUE);
[13]       myseed = as.numeric(args[1]);
[14]       design = as.numeric(args[2]);        # design: 1-2
[15]       car.design = as.numeric(args[3]);    # design: 1-2
[16]       G = as.numeric(args[4]);             # number of clusters
[17]       Nmax = as.numeric(args[5]);          # Maximum number of possible obs within cl
[18]  # myseed=1111
[19]  # design=2;
[20]  # car.design=1;
[21]  # G=5000;
[22]  # Nmax=500;
[23]       MC=5000;
[24]       alpha=0.05;
[25]       mu.1=0;
```

To obtain exactly the same results as those in the paper, you should comment out lines [12]-[17] and then undo the comment out of lines [18] to [22] and replace the values with those in the following files listed next.

3. `param-NICS.txt`, `param-NICSv2.txt`, `param-NICSv3.txt`, and `param-NICSv4.txt`. The format in those files looks like

```
1011020 1 1 1000 200
2011020 2 1 1000 200
1011050 1 1 1000 500
2011050 2 1 1000 500
....
....
```

where each row is a simulation run and each column contains: the first column contains the seed (args[1]), the second column contains the design (args[2]), the third column contains the CAR design (args[3]), the fourth column contains the number of clusters (args[4]), and the last column contains the parameter Nmax that is defined in the paper (args[5]).

The output of `NICS.Main.R` is either a LaTeX table or a txt output with an easy to visualize table. Both of these outputs are placed in independent subfolders that will be created when running the code.

4. `CARArrayscript.sh`: contains an example script for illustration purposes only.

Readme file for empiricalApplication.m

Created by Federico A. Bugni (federico.bugni@northwestern.edu), April 2024.

Description

This MATLAB program is designed to implement the empirical illustration described in Section 5 of the paper "Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes" by Bugni, Canay, Shaikh, and Tabord-Meehan (henceforth BCST24), accepted for publication in JPE Micro in 2024. It processes data to produce the figures and tables in Section 5 of BCST24.

Inputs

The program requires a data file named empiricalApplicationData.mat which should be placed in the root directory of the project.

This data file originated from maindata.dta from the replication file by Celay et al. (2019, AEJ:applied). It consists of 3 matrices corresponding to different periods: pre-intervention (data1.mat), intervention (data2.mat), and post-intervention (data3.mat).

Each matrix is has a list of patient-level observations (i=1,...,N_g) for each clinic (i.e., clusters) (g=1,...,G) organized in 5 columns:

- Column 1: the week of the first prenatal visit (Y_{i,g} outcome 1),
- Column 2: whether the visit was before the 13th week (Y_{i,g} outcome 2),
- Column 3: treatment assignment of the clinic (A_g),
- Column 4: identity of the clinic (g),
- Column 5: number of observations per clinic (N_g).

Outputs

The MATLAB script generates all the material in Section 5 of BCST24:

- Figure 2: Histogram of the week of the first prenatal visit during the pre-intervention period.
- Figure 3: Histogram of patients per clinic during the intervention period.
- Table 7: Estimation results based on the structured data.

Usage

1. Place the empiricalApplication.m and empiricalApplicationData.mat in the same directory.
2. Run the script empiricalApplication.m in MATLAB. The script will automatically load the necessary data, perform the analysis, and save the outputs in the specified format within the directory.