Replication Codes and Data

"Random Assignment with Non-Random Peers: A Structural Approach to Counterfactual Treatment Assessment"

Review of Economics and Statistics

Alan Griffith
University of Washington

March 29, 2022

Overview

There are two main sets of analysis, which are discussed in turn below. The reduced-form analysis is performed in Stata, while the structural analysis is done in Matlab. *Both* parts are necessary to fully replicate the paper's analysis. If you require additional information or find errors, please contact at alangrif@uw.edu.

1 Reduced-form Analysis in Stata

(Final paper analysis run in Stata/SE 15.1.)

Inputs

1. Stata .do file
   (a) **Griffith REStat Stata Mar2022.do** – main analysis file that performs the entire reduced-form analysis
2. Stata datasets

(a) **Indiv_Analysis_imp_Jun2021.dta** – individual-level data on outcomes and covariates. This data is unique by (schl_code enroll_num).

(b) **Linkwise_Analysis_Sample_Jun2021.dta** – link-level data on continuous network links. This data is unique by (schl_code enroll_num_self enroll_num_alter)

(c) **IDs_and_levels.dta** – dataset that helps to merge the other two for purposes of performing the analysis. This data is unique by (schl_code enroll_num).

Outputs

(Table numbers correspond to tables in final paper and supplemental material.)

1. **Table1.tex**
2. **Table2.tex**
3. **Table3.tex**
4. **Table4.tex**
5. **Table5.tex**
6. **TableA5.tex**
7. **TableA6.tex**
8. **TableA7.tex**
9. **TableA8.tex**
10. **TableA9.tex**
11. **TableA10.tex**
12. **TableA11.tex**
13. **TableA12.tex**

Instructions

Place all files into the working directory and run the do file. After analysis is run, the output tables will appear in the working directory.

2 Structural Analysis in Matlab

(Final paper analysis run in MATLAB R2018b.)

Inputs

1. Matlab code files (in order of when first used in estimation)

   (a) **Griffith_REStat_Matlab_Mar2022.m** – runs the entire analysis. See Notes below for discussion of settings.

   (b) **CreateMats_Mar2022.m** – creates data matrices from raw data.

   (c) **Network_EM_Mar2022.m** – runs EM algorithm to estimate structural network model.

   (d) **Fn_EstimateNetwork_Mar2022.m** – function that estimates the network, running inside *Network_EM_Mar2022.m*.

   (e) **Fn_ImputeNetwork_Mar2022.m** – function that imputes the network, running inside *Network_EM_Mar2022.m*.

   (f) **GMM_Fn_Mar2022.m** – GMM criterion for estimation of the network model, called in *Fn_EstimateNetwork_Mar2022.m*. Also used in estimating variance, called in *Variance_EM_Mar2022.m*.

   (g) **parload_b.m** – function to load estimates as part of parallel loops of network EM algorithm, used inside *Network_EM_Mar2022.m*.

   (h) **parsave_b.m** – function to save estimates as part of parallel loops of network EM algorithm, used inside *Network_EM_Mar2022.m*.

   (i) **PE_EM_Mar2022.m** – takes estimates of structural network model and runs EM algorithm to estimate structural peer effects parameters.

   (j) **Fn_EstimatePE_Mar2022.m** – function that estimates the peer effects model, running inside *PE_EM_Mar2022.m*.

   (k) **Fn_ImputePE_Mar2022.m** – function that imputes outcomes via the peer effects model, running inside *PE_EM_Mar2022.m*.

   (l) **Estimate_NoA_Mar2022.m** – runs estimation of peer effects model without scalar unobservables.

   (m) **Variance_EM_Mar2022.m** – takes structural estimates of network and peer effects parameters from EM algorithm and estimates variance.

(n) **Run_Pref_Sims_Mar2022.m** – takes estimtes of structural model and simulates outcomes under preferred assignments.

(o) **Sim_Mar2022.m** – takes estimates of structural model and simulates outcomes. Also used inside *Run_Pref_Sims_Mar2022.m*.

(p) **Tables_Mar2022.m** – takes estimates and populates tables with structural parameters.

(q) **matrix2latex.m** – function that creates .tex output from matrices.

2. Data in .csv format

(a) **Pairwise_T2C.csv** – pairwise network data for 20 schools comprising Random Treatment and Control. Columns 1 and 2 comprise the continuous network measure and an indicator for where this measure is missing in the raw data, respectively.

(b) **Pairwise_T1.csv** – pairwise network data for 10 schools comprising Elected Treatment. Columns 1 and 2 comprise the continuous network measure and an indicator for where this measure is missing in the raw data, respectively.

(c) **Indiv_T2C.csv** – individual-level covariate and outcome data for 20 schools comprising Random Treatment and Control. Col. 1-7 are indicators for Participant, Elected, Standard 7, Standard 8, SC, ST, and OBC. Cols. 8-10 are indicators for predicted tercile (L, M, H) for Educational Aspirations. Cols. 11-13 are indicators for predicted tercile (L, M, H) for Gender Roles. Cols 14-15 are endline Educational Aspirations and Gender Roles indices, respectively. Cols 16-17 are baseline Educational Aspirations and Gender Roles indices, respectively. Cols 18-19 are indicators for missing values of endline Educational Aspirations and Gender Roles indices, respectively. Cols 20-21 are indicators for missing values of baseline Educational Aspirations and Gender Roles indices, respectively.

(d) **Indiv_T1.csv** – individual-level covariate and outcome data for 10 schools comprising Elected Treatment. Col. 1-7 are indicators for Participant, Elected, Standard 7, Standard 8, SC, ST, and OBC. Cols. 8-10 are indicators for predicted tercile (L, M, H) for Educational Aspirations. Cols. 11-13 are indicators for predicted tercile (L, M, H) for Gender Roles. Cols 14-15 are endline Educational Aspirations and Gender Roles indices, respectively. Cols 16-17 are baseline Educational Aspirations and Gender Roles indices, respectively. Cols 18-19 are indicators for missing values of endline Educational Aspirations and Gender Roles

indices, respectively. Cols 20-21 are indicators for missing values of baseline Educational Aspirations and Gender Roles indices, respectively.

(e) **sch_size_T2C.csv** – school-level data on the number of students in each school for 20 schools comprising Random Treatment and Control.

(f) **sch_size_T1.csv** – school-level data on the number of students in each school for 10 schools comprising Elected Treatment.

Outputs

(Table numbers correspond to tables in final paper and supplemental material.)

1. **Table6.tex**
2. **Table7.tex**
3. **TableA3.tex**
4. **TableA4.tex**
5. **TableA15.tex**
6. **TableA17.tex**
7. **TableA19.tex**

Instructions

Place all files into the working directory and run Griffith_REStat_Matlab_Mar2022.m. After analysis is run, the output tables will appear in the working directory.

Notes

1. In order to keep the code as simple as possible, I have omitted the analyses that perform the following two robustness checks:

   (a) Alternative Network Definition (results in Appendix Tables A.13, A.14, A.16, and A.18)

   (b) Sensitivity Analysis discussed in Appendix C.3 (results in Appendix Table A.1)

2. The following settings are adjustable. They should be set in the preamble of the main analysis file *Griffith_REStat_Matlab_Mar2022.m*.

(a) *Restart* should be set to 0 initially. If the network EM algorithm freezes or shuts down for some reason, change Restart to 1, then re-initiate.

(b) *chains* sets the number of parallel chains for the network EM algorithm. This was set to 2 for the paper's analysis.

(c) *sim_tol* sets the optimization tolerance for the GMM criterion (using fmincon) for the network estimation step at each iteration of the network EM algorithm. This was set at 1e-6 for the paper's analysis.

(d) *reps_chain* sets the number of repetitions for each parallel chain for the network EM algorithm. This was set at 500 for the paper's analysis.

(e) *PE_sim_reps* sets the number of iterations for each run of the peer effects EM algorithm. This was set at 500 for the paper's analysis.

(f) *burnin* sets the "burn-in" period before taking draws from the chain of the EM algorithm. This was set at 100 for the paper's analysis.

(g) *b_gap* sets the gap between draws from the chain of the network EM algorithm. This was set at 20 for the paper's analysis.

(h) *chains_NoA* sets the number of parallel chains for the EM algorithm that estimates the model without scalar unobservables. This was set at 500 for the paper's analysis.

(i) *Sim_reps* sets the number of simulation repetitions for simulations for purposes of comparing to Elected Treatment and simulating preferred assignments. This was set at 1000 for the paper's analysis.