Description of supplementary material for "Random Coefficients on Endogenous Variables in Simultaneous Equations Models"

Matthew A. Masten
Department of Economics
Duke University
matt.masten@duke.edu

July 14, 2017

1 Supplementary online appendix

The supplemental online appendix provides several results which complement those in the main paper. Among other things, it includes Monte Carlo simulations to examine the finite sample performance of the proposed estimator. I discuss the code used for these simulations in section 4 below.

2 Data used in the empirical illustration

As cited in the acknowledgments of the main paper, this paper uses data from two studies: (1) The National Longitudinal Study of Adolescent to Adult Health ("Add Health") and (2) The Adolescent Health and Academic Achievement (AHAA) study. The Add Health study is well known for its collection of network data, among other features. Because I rely on this network data to construct friend pairs, I use the restricted-use version of the data. Information on how to obtain this data is available on the Add Health website (http://www.cpc.unc.edu/addhealth). I next describe specifically what data must be requested in order to replicate the results in this paper.

There are two basic parts to the Add Health data: (1) the main restricted use data set, which is provided to all researchers who requests access to the restricted use data, and (2) several constructed datasets. Each constructed dataset must be specifically requested from Add Health. Importantly, the data to construct friend links is only contained in a constructed dataset. Among all the data available from Add Health, in this paper I only use the following:

1. Main restricted use dataset
    - In-home interview files

  - Wave 1 (formerly ICPSR study #27021/DS0001).
- School files
  - In-school questionnaire (from wave 1) (formerly ICPSR study #27021/DS0019).

2. Constructed datasets
   - Friend Files
     - Wave 1 in-home friend nominations (formerly ICPSR study #27022/DS0001). This dataset is used to construct best friend pairs.
     - Wave 1 in-school friend nominations (formerly ICPSR study #27022/DS0003). This dataset is used in my code, but is not used in the final analysis. Hence one could omit requesting this data by modifying the code appropriately.
   - Education Files
     - Academic courses (formerly ICPSR study #27030/DS0001). This dataset contains student GPAs, from the AHAA study.

Further documentation and codebooks for these datasets are publicly available on the Add Health website.

3 Stata code for the empirical illustration

I used Stata/SE version 13.1 on MacOS 10.9.5 to process the raw data for the empirical illustration, compute the summary statistics of table 1, and compute the standard estimators reported in tables 2 and 3. There are two folders containing Stata code:

1. `cr_data/programs`. This folder contains all the code to process the raw data into the final datasets used in the analysis. The file `master.do` runs all of the code in sequence. To run that file you first must specify the file paths containing the data and the code.

   One important point is that I obtained the Add Health data when it was administered by the Inter-university Consortium for Political and Social Research (ICPSR). In the summer of 2014, management of this data was transferred to the Carolina Population Center (CPC) at the University of North Carolina at Chapel Hill (UNC-CH). My code is written for the ICPSR provided versions of the datasets, and hence the raw data file names and directories may have since changed. Hence you may have to update these references throughout the code to match your version of the data.

2. `analysis/programs`. This folder contains the code to produce some baseline empirical results. `master_analysis.do` runs all of the Stata analysis code. Again, you must first specify the file paths containing the data and the code. `summaryStats.do` produces table 1.

`main_IV_regs_3SLS.do` produces results for table 2 and the first two columns of table 3. Unlike table 1, tables 2 and 3 are not automatically produced from the code and their entries must be manually copied.

4 Matlab code for simulations and the empirical illustration

I used Matlab to implement the estimator described in section 4 of the paper, for the empirical illustration and the Monte Carlo simulations in the supplementary online appendix. Here I describe how to use these files to reproduce these results; further details on the implementation are provided as comments in the source code.

Monte Carlo simulation

Because running the simulations linearly on a single machine would take far too long, I used the Duke Computing Cluster to run the code for each draw of the data in parallel. For this reason, there is not a single master file which can reproduce all of the simulation results in a single run. Instead, below I outline how to use this code with a computing cluster to reproduce the results. I also again thank Margaux Luflade for her excellent work solving many of the programming challenges for this research.

1. Run `runGenerateDatasets.m` on the cluster to generate all the Monte Carlo datasets.
2. For each dgp, run `runComputeMISE_cluster.m` via a batch job array on the cluster. The array index selects a value of `radonBandwidth`. The total number of jobs is specified by the length of `Estimationparams.radonBandwidths` in `setRadonBandwidths.m`.
3. Download MISE datasets to your local computer.
4. Run `computeOptimalBandwidthsFromCluster.m` on your local computer to compute optimal bandwidths.
5. Upload optimal bandwidth datasets to the cluster.
6. For each simulation dataset, run `runWithEstimatedFirstStage_cluster.m` on the cluster.
7. Download simulation estimates to your local computer.
8. Run `runTSLSestimationWrapper.m` to compute 2SLS estimates for all simulation datasets.
9. Run `runComputeMCresults.m` to aggregate simulation results and compute all desired Monte Carlo results reported in the paper. This will also produce figures 2 and 3 of the supplementary online appendix.
10. Run `printMCresults.m` with inputs 1 and 2 to produce tables 1 and 2 of the supplemental online appendix.

Other remarks:

- Use `plotLinearCombPdf.m` to produce figure 4 of the supplementary online appendix.
- `runEmpiricallyCalibratedDgpStats.m` computes the average first stage F-statistic and average mass near $[0, 4]$ for the empirically calibrated dgp.
- The Matlab folder contains the entire file structure used by the code. These empty folders are filled once the above code is run.

Finally, for all results I used Matlab version R2014a locally, on Mac OS 10.9.5. The Duke Computing Cluster used Matlab version R2012b.

Empirical illustration

The following remarks explain how to reproduce the empirical illustration results.

- Run `runEmpiricalAnalysisWithAllBandwidths.m` to produce all of the main empirical results: column 3 of table 3 of and figure 1 of the main paper, as well as tables 3 and 4, and figure 5 of the supplemental online appendix. This can be done on a local computer. As with the Stata code, you need to change the file path in `runEmpiricalAnalysis.m` to the location of the dataset outputted by Stata. You also must first run the Stata code described in section 3 above to obtain the final dataset used in this Matlab code.
- `runEmpiricalAnalysis.m` contains code (commented out) for producing the numbers in supplemental online section appendix B.4.