README

Paper: Omitted variable bias of Lasso-based inference methods: A finite sample analysis

Authors: Kaspar Wuthrich (kwuthrich@ucsd.edu) and Ying Zhu (yiz012@ucsd.edu)

Software

Hardware and operating system

The code was run on a MacBook Pro (2020) and Mac mini (M1, 2020) with macOS Big Sur (Version 11.5.2).

Statistical software

- R version 4.1.1 (2021-08-10) -- "Kick Things" (x86_64-apple-darwin17.0 (64-bit))
- Matlab R2020b (9.9.0.1467703 64-bit (maci64))
- Stata 17.0 MP-Parallel Edition

List of files

Code for data preparation for the empirical studies

- `data_prep_401k.m:` Data preparation for the empirical study of the effect of 401(k) plans on savings in Section 5.
- `data_prep_frylerlevitt.do:` Data preparation for the empirical study of the racial differences in the mental ability of children in Section 5.

Generic functions

- `pdl_opt_lambda_sel_prob_tau.m:` Post double Lasso implementation based on the regularization choice in Bickel et al. (2009).
- `gen_design_r2.m`: Code to generate data for the numerical example in Section 3 and C.1.

- `noncollinear_Revision.m`: Function used by `data_prep_401k.m` from the replication package accompanying Belloni et al. (2017, ECMA), downloaded from the Econometrica website:

  Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Supplement to "Program evaluation and causal inference with high-dimensional data". Econometrica Supplemental Materials. URL: https://www.econometricsociety.org/publications/econometrica/2017/01/01/program-evaluation-and-causal-inference-high-dimensional-data (last accessed 09/06/2021)

- `Rfunctions.R`: Generic R codes for generating data, post double Lasso with CV, OLS with robust and HCK standard errors, and debiased Lasso.

Code for replicating the figures and tables in the paper

- `intro_graph.m`: Code to generate Figure 1 in the main text.

- `example_distribution.m`: Code to generate the figures based on the numerical example. Generates Figures 2 and 3 in the main text and Figure C.1 in the appendix.

- `simulations.R`: Code for the simulation studies in the main text and the appendix. Generates Figures 4-7 and 10-11 in the main text and Figures B.1, C.2, C.3, and C.4 in the appendix.

- `app_401k.R`: Code for replicating the results of the empirical study of the effect of 401(k) plans on savings in Section 5. Generates Table 1 and Figure 8. Requires running `data_prep_401k.m` first.

- `app_fryerlevitt.R`: Code for replicating the results of the empirical study of the racial differences in the mental ability of children in Section 5. Generates Table 2 and Figure 9. Requires running `data_prep_frylerlevitt.do` first.

Data

- `restatw.dat`: Data from the 1991 SIPP used in Belloni et al. (2017, ECMA), downloaded from the Econometrica website:

  Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. (2017). Supplement to "Program evaluation and causal inference with high-dimensional data". Econometrica Supplemental Materials. URL: https://www.econometricsociety.org/publications/econometrica/2017/01/01/program-evaluation-and-causal-inference-high-dimensional-data (last accessed 09/06/2021)

  The main variables of interest are total wealth (*tw*) and an indicator for 401(k) eligibility (*e401*). The dataset also contains rich additional information on the households; see Belloni et al. (2017, ECMA) for a detailed description.

- `cpp.dta:` Data from the US Collaborative Perinatal Project used in Fryer & Levitt (2013, AER), downloaded from public replication archive on ICPSR:

  Fryer, Roland G., J. and Levitt, S. D. (2013). Replication data for: Testing for racial differences in the mental ability of young children. Nashville, TN: American Economic Association [publisher], 2013. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-11. https://doi.org/10.3886/E112609V1 (last accessed 09/06/2021)

  The main variables of interest are IQ at the age of seven (*stand_fullscale_iq_7years*) and an indicator for Black children (*black*). The control variables include extensive information on socio-demographic characteristics, the home environment, and the prenatal environment; see Table 1B in Fryer & Levitt (2013, AER) for a detailed description.

  The replication package also includes the original license file from ICPSR named `ORIGINAL_LICENSE_FRYER_LEVITT.txt`.

Instructions

Running the Stata, Matlab, and R-codes requires putting all files into the same folder and adjusting the working directory in each file accordingly. Running the R-codes requires installing the relevant packages first.