

Status Summary

Lucas Cruz Fernandez

01.09.2021

Overview

Approach

Problems, Ideas, Questions

Econometric Model

- ▶ want to estimate the Partially Linear Model

$$Y_{it} = \theta(X_{it})D_{it} + g(X_{it}, W_{it}) + \epsilon_{it} \quad (1)$$

$$D_{it} = h(X_{it}, W_{it}) + u_{it} \quad (2)$$

- ▶ $\theta(X_{it})$ is the constant marginal CATE
- ▶ this allows calculating MPC for each household in each period based on their X_{it}

Estimator

- ▶ original DML only reasonable for cross-section
- ▶ JPS and MS also only use cross-section estimators
 - implicitly assume strict exogeneity, which is unreasonable
- ▶ use Panel Double Machine Learning Estimator (DML) by Chernozhukov et al. (2021)
- ▶ since treatment dimension fixed ($d_T = 1$) only difference to DML is first stage cross-fitting algorithm
 - see Section 3.1 of Chernozhukov et al. (2021)

Estimator

- rewrite the PLM to

$$Y_{it} - E[Y_{it}|X_{it}, W_{it}] = \theta(X)(D_{it} - E[D_{it}|X_{it}, W_{it}]) + \epsilon_{it}$$

- use a machine learning estimator to find estimators for

$$d_{i0} = E[Y_{it}|X_{it}, W_{it}]$$

$$l_{i0} = E[D_{it}|X_{it}, W_{it}] (= h(X_{it}, W_{it}))$$

- get respective residuals and estimate $\theta(X)$ using OLS

Algorithm 1 Panel DML Recipe

- 1: Partition the data into K -folds based on their time index.
An observation is added to partition I_k if:

$$I_k = \{(i, t) : \lfloor T(k-1)/K \rfloor + 1 \leq t \leq \lfloor Tk/K \rfloor\}$$

- 2: For each partition k use a first stage estimator to estimate \hat{d}_k and \hat{l}_k using data of all folds except k (cross-fitting).
 - 3: Orthogonalize treatment and outcome of observation (i, t) using the predictions of the corresponding fold to get the residuals.
 - 4: Use all residuals to estimate the coefficient θ using a suitable estimator.
-

Panel DML

- ▶ Panel DML needs lag structure but only have at most 3 observations of i
- ▶ when only using i that have $T = 3$ sample size reduced drastically
- ▶ one-period lag should be sufficient as looking at quarterly data, hence two periods for each household in reduced dataset
- ▶ idea: run two (three) specifications
 1. without lags and larger N
 2. with lags and smaller N
- ▶ compare whether effects differ strongly
- ▶ advantage: Panel DML actually captures these dynamics compared to approaches by JPS and MS

What are X ?

- ▶ choosing $X = Z$ leads to undertermined covariance matrix of the estimator
- ▶ hence, what variables should be X ?
- ▶ data-driven way not feasible as this could change derived inference results
- ▶ use economic reasoning but what is reasoning then?
Could find something for every variable I guess...

Single variable effects

- ▶ estimate the constant marginal CATE $E[Y_1 - Y_0|X = x]$
- ▶ X is a vector of variables, hence cannot see single variable effects on treatment effect
- ▶ idea 1: simply look at correlations between point estimates and variable
- ▶ idea 2: use the Marginal Effect at the Means, i.e. setting all $X_{\setminus j}$ at their cross-sectional mean to get X_j 's marginal CATE at the mean
 - major issue: depends extremely on test/train sample split, no matter what way is chosen to train model
 - why? salary and total family income have a giant variance