Correcting Overfitting in IV regression: an immigration example

Research Question: How can we use machine learning methods to correct overfitting in IV regression?

Haowen Shang, MACSS Project Proposal, Spring2019

Motivation

- Over half of the foreign-born doctorates remain in the United States (Michael Finn, 2003) , suggesting they may have a sizable impact on the labor market for high-skill workers.

- More Strict Visa application for doctoral student (Administrative Processing Back Ground Check)

- How do foreign students affect high skilled labor market?

Motivation

- ➢Borjas, George J. "The Labor-Market Impact Of High-Skill Immigration," American Economic Review, 2005, v95(2,May), 56-60
- a foreign student influx into a particular doctoral field at a particular time had a significant and adverse effect on the earnings of doctorates in that field who graduated at roughly the same time.
- A 10 percent immigration-induced increase in the supply of doctorates lowers the wage of competing workers by about 3 percent.

Model

- ➢ $logw_{ifc}(t) = \ v_{ifc} + x_{ifc}(t) + \pi_t + \left(d_f * \pi_t\right) + \varepsilon_{ifc}(\mathrm{t})$
- ➢ $\widehat{v_{fc}} = \eta logL_{fc} \ + \ d_f + y_c + \xi_{fc}$

- w_{ifc} is the is the the annual earnings of worker i, who has a doctorate in field f, received his doctoral degree in year c, and is observed at time t.
- v_{ifc} is the individual fixed effect.
- $x_{ifc}(t)$ is a vector indicating the number of years that the worker has been in the labor market.
- d_f is a vector of fixed effects indicating the worker’ s field of doctoral study.
- π_t is a vector of period fixed effects indicating the calendar year in which the worker’ s earnings are observed.
- $\widehat{v_{fc}}$: we use the total of the sampling weights assigned to each person in the SDR calculate the average v_{fc}
- L_{fc} is the total number of foreign doctorates in field f and cohort c
- y_c is a vector of fixed effects indicating the worker’ s year-of-graduation cohort.

Contributions

- $logw_{ifc}(t) = \ v_{ifc} + x_{ifc}(t) + \pi_t + \left(d_f * \pi_t\right) + \varepsilon_{ifc}(\mathrm{t})$
- $\widehat{v_{fc}} = \eta logL_{fc} + \ d_f + y_c + \xi_{fc}$

- full-sample estimation may have overfitting problems
- Want to use bootstrapping method to correct overfitting of this IV model

Data

➢the Survey of Earned Doctorates

- The SED provides a *population* census of all doctorates granted by U.S. institutions, with a response rate of around 92 percent.
- We use the SED to calculate the magnitude of the immigrant supply shock by field and year of degree(L_{fc}).

➢the Survey of Doctoral Recipients

- The SDR is a biennial longitudinal file that provides a 7 percent sample of doctorates in science or engineering granted by U.S. institutions, and contains detailed information on a worker’s earnings.

Method and Procedure

- 1. run Borja(2005) regression on the full dataset and got the parameters $\hat{\beta}^{Bor}$

- 2. estimate $\hat{\beta}^{New}$ using bootstrapping

- 3.Compare the MSE of the model with $\hat{\beta}^{Bor}$ and $\hat{\beta}^{New}$

- 4.Discuss the differences of $\hat{\beta}^{Bor}$ and $\hat{\beta}^{New}$

Method and Procedure

- $\hat{\beta}^{New} = \frac{1}{N}\sum_{s=1}^{N}\hat{\beta}^{s}$,where $\hat{\beta}^{s}$ is the estimated parameter of each training set

- $\widehat{MSE}^{new} = \frac{1}{N}\sum_{s=1}^{N} MSE^{s,new}$, where $MSE^{s,new}$ is the MSE from each random test set

- $\widehat{MSE}^{Bor} = \frac{1}{N}\sum_{s=1}^{N} MSE^{s,Bor}$

Method and Procedure

- $\hat{\beta}^{New} = \frac{1}{N}\sum_{s=1}^{N}\hat{\beta}^{s}$,where $\hat{\beta}^{s}$ is the estimated parameter of each training set

- $\widehat{MSE}^{new} = \frac{1}{N}\sum_{s=1}^{N}MSE^{s,new}$, where $MSE^{s,new}$ is the MSE from each random test set

- $\widehat{MSE}^{Bor} = \frac{1}{N}\sum_{s=1}^{N}MSE^{s,Bor}$

Potential Outcomes and Future Works

- $\widehat{MSE}^{new} < \widehat{MSE}^{Bor}$

- Expand this method into other regression examples
- Using other methods to correct the overfitting problems