Correct Overfitting in Two Stage OLS Regression: an Immigration Example

Haowen Shang (haowen@uchicago.edu)
Division of the Social Sciences, University of Chicago

MASTERS IN COMPUTATIONAL SOCIAL SCIENCE
THE UNIVERSITY OF CHICAGO

Research Question

Can we use machine learning methods to correct overfitting problems in two stage OLS regression?

Related Literature

My work is based on the efforts of Borjas on exploration of high skilled labor markets. The empirical results showed that "The influx of foreign students has a significant and adverse effect on the earnings of competing doctorates. A 10 percent immigration-induced increase in the supply of doctorates lowers the wage of competing workers by about 3 to 4 percent." This was a very strong conclusion. However, it may have overfitting problems because Borjas used full sample as training set to do the estimation, while maximizing accuracy in training set can pick incorporate noise from the data.

Contribution to the Literature

I want to detect and correct the overfitting problems in two stage OLS regression and find more accurate and powerful estimation parameters.
I will use Borjas' immigration model to do the two stage OLS regression and explore the impact of foreign student on high skilled labor market. The immigration shock is measured by the percentage of foreign student and the high skilled labor market outcomes are measured by the annual earnings of doctoral recipients.

Procedures:

- Use full sample to do the estimation.
- Use bootstrap method to do the estimation.
- Compare the mean squared error of these two estimations,

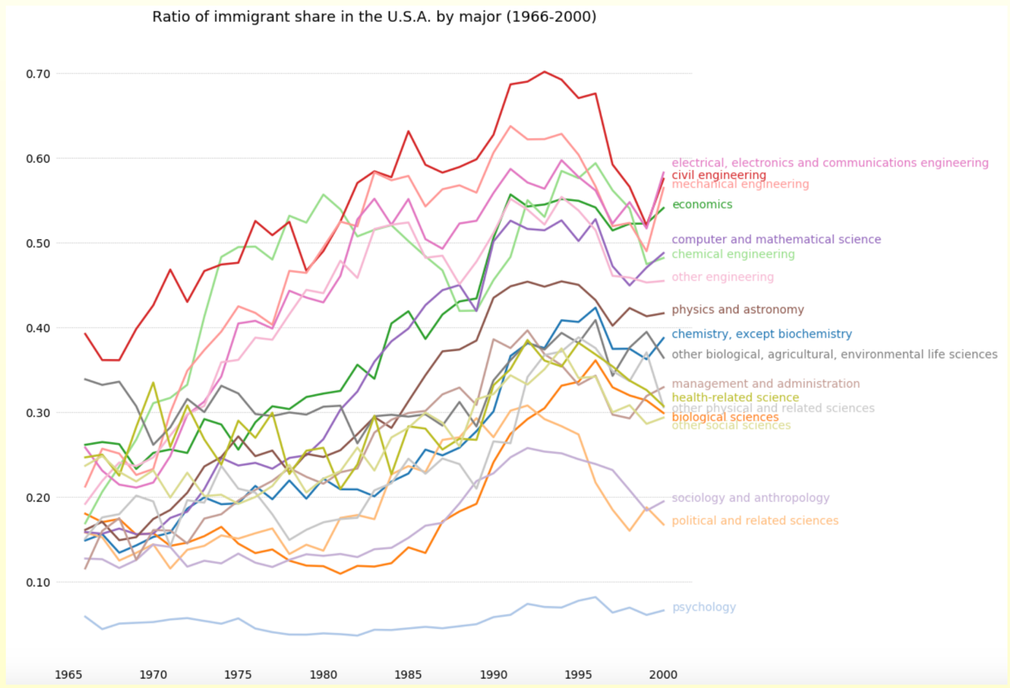
Data Sources

- Survey of Earned Doctorates (SED)
- Survey of Doctoral Recipients (SDR)

Table 1: Summary Statistics of Key Variables

variables	Annual Earnings	Immigrant Share
Count	92078	92078
Mean	94150.731	0.288
Std	40833.232	0.164
Min	0.000	0.041
25%	65000	0.154
50%	95000	0.293
75%	130000	0.396
Max	150000	0.702

Data Description



Model

- The immigration model

$$\log W_{ifc}(t) = \theta\, p_{fc} + x_{ifc}(t) + d_f + y_c + \pi_t + \varepsilon_{ifc}(t)$$

However, since one individual can appear several times in the database in different SER conducted year and immigrant share does not change for individuals who graduated in same year and studied in same field, so the error terms for OLS regression will be incorrect. Borjas handled this model into a two stage OLS regression to make the error terms more correct:

$$\log W_{ifc}(t) = V_{ifc} + x_{ifc}(t) + \pi_t + \varepsilon_{ifc}(t)$$

$$\widehat{V_{fc}} = \theta p_{fc} + d_f + y_c + w_{fc}$$

Model

- The bootstrap model

Firstly, I will use Borjas' model and method to do this two stage OLS regression. He uses the full sample to do this regression. We can get the estimation parameter $\hat{\beta}^{FS}$ and mean squared error MSE^{FS} from this full sample regression.

$$MSE^{FS} = \frac{1}{n}\sum_{i=1}^{n}(Y_i^{FS} - \hat{Y}_i^{FS})^2$$

Then I will use bootstrap method to do this regression. I will randomly split the whole data set into training set and testing set. Then use the training set to get all the estimation parameters and use these parameters to calculate the MSE.

$$MSE^{BS} = \frac{1}{n}\sum_{i=1}^{n}(Y_i^{BS} - \hat{Y}_i^{BS})^2$$

Then repeat this procedure N times and calculate the average mean squared error of regression using full sample estimation parameters and using bootstrap estimation parameters.

$$\overline{MSE} = \frac{1}{N}\sum_{i=1}^{M} MSE$$

Finally, compare $\overline{MSE}^{FS}$ and $\overline{MSE}^{BS}$ to find whether bootstrap method can make the model better fit. If we find that:

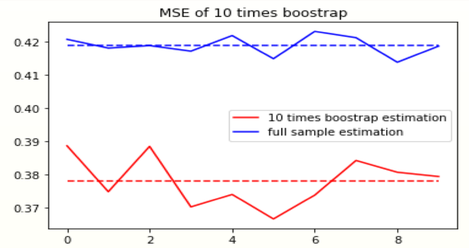
$$\overline{MSE}^{BS} < \overline{MSE}^{FS}$$

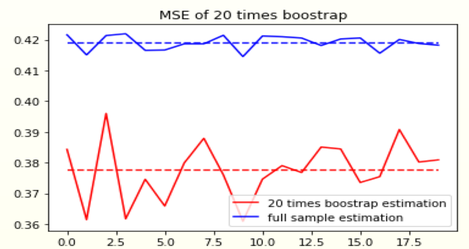
we can have a conclusion that bootstrapping can detect and correct overfitting problems in this two stage OLS model.

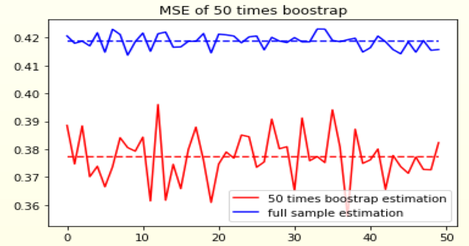
Reference

Borjas, George J. "The labor-market impact of high-skill immigration." *American Economic Review* 95, no. 2 (2005): 56-60.

Results







	Full Sample	10 times bootstrap	20 times bootstrap	50 times bootstrap
Immigrant share coefficient	0.2296	0.2286	0.2271	0.2243
MSE(FS)	0.4309	0.4187	0.4190	0.4186
MSE(BS)		0.3780	0.3775	0.3772

Conclusion

Bootstrap estimation has smaller mean squared error than full sample estimation. Increasing bootstrap times doesn't influence mean squared error so much. θ is 0.224 in 50 times bootstrap. We know that the more accurate impact of foreign students on high skilled labor market is that a 10% increase of immigrant will decrease native's earnings by approximately 1.1%.