Correct Overfitting in Two Stage OLS Regression: an Immigration Example *

Haowen Shang †

June 12, 2019

Abstract

Today, many researchers use full dataset to do the estimation, while full sample estimation may pick incorporate noise from the dataset and have overfitting problems. This research wants to use machine learning methods to correct the overfitting problem in a two stage OLS regression and find more accurate and powerful parameters. This research picks an immigration model to do the two stage OLS regression and explore the impact of foreign student on high skilled labor market outcomes. Firstly, use full sample to do the regression and then use bootstrap method to do the regression again. By comparing the mean squared error of these two estimations, this paper finds that the mean squared error of bootstrap regression is smaller than the full sample regression, which means the parameters got from bootstrap method are more accurate and have better prediction power.

keywords: full sample, bootstrap, overfitting, mean squared error, immigration.

*I would like to express my thanks to academic instructions from Professor Richard Evans, the director of Master in Computational Social Science program at the University of Chicago.

†University of Chicago, Division of the Social Sciences, Master of Computational Social Science, haowen@uchicago.edu.

1 Introduction

Researchers always want to find the model that best describes the data. They may use the full sample dataset to find the model with best accuracy, but this model may not perform well when used into other out of sample dataset. Silver (2012) pointed out that researchers want to estimate based on data's 'signal' rather than 'noise'. However, when they just use full sample dataset as training set to get the estimation model, the model may capture noise instead of finding the signal. It will then just perform well on its training data but can not be generalized into other datasets. Hawkins (2004) pointed out that this is the problem of overfitting.

How can we correct overfitting problems of full sample estimation? A resampling method named bootstrap is useful for correcting overfitting problems. Steyerberg (2009) illustrated that bootstrap resampling was a central technique to correct overfitting and quantify optimism in model performance . James et al. (2013) introduced that bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method. It's difficult for us to get new sample from whole underlying population. We can just obtain a small part of data from the population. The bootstrap approach allows us to use a computer to emulate the process of obtaining new sample sets, so that we can estimate the variability of our model without generating additional samples. Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations from the original data set.

This paper picks an immigration model created by Borjas (2005) and Borjas (2009) to find whether bootstrap method can correct overfitting problems. Specifically, the model is a two stage OLS estimation model that explore the impact of immigration on high-skilled labor market outcomes. This model considered the uneven distribution of immigrants across local labor markets and flows of internal migrants and took skill group and experience into consideration. By using panel data of doctorates recipients and instrument variable regression, this model can estimate the impact

of supply shock by foreign doctoral students on annual earnings of native doctoral students. The empirical results in Borjas (2009) showed that "The influx of foreign students into a particular field at a particular time has a significant and adverse effect on the earnings of competing 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 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. What this paper wants to contribute to Borjas' research is to correct the overfitting problem in this two stage OLS regression and find better estimation parameters. I will use bootstrap method to redo the regression model created by Borjas (2005) and Borjas (2009), get new estimation parameters and find whether the prediction power would be improved by using this method.

2 data

2.1 Data Source

The data used in this research are from two surveys. One is the 'Survey of Earned Doctorates (SED)'. The other is the 'Survey of Doctoral Recipients (SDR)'.

The Survey of Earned Doctorates (SED) is an annual census conducted since 1957 of all individuals receiving a research doctorate from an accredited U.S. institution in a given academic year. The SED collects information on the doctoral recipient's educational history, demographic characteristics, and postgraduation plans. Results are used to assess characteristics of the doctoral population and trends in doctoral education and degrees. SED data is available until 2017, but I will just use SED data from 1966 to 2000, because the major categories in this dataset changed since 2000 and there are many missing data after 2000.

The Survey of Doctorate Recipients (SDR) provides demographic, education, and

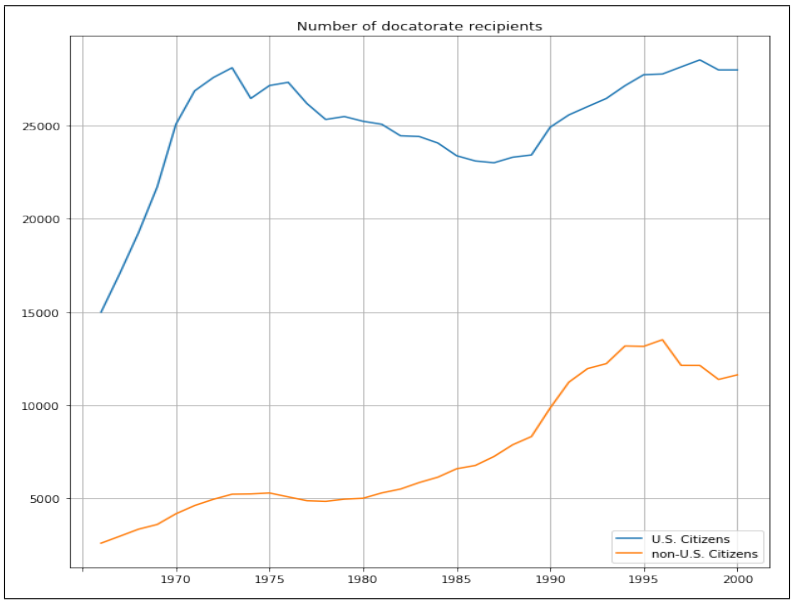
career history information from individuals with a U.S. research doctoral degree in a science, engineering, or health (SEH) field. Conducted since 1973, the SDR is a unique source of information about the educational and occupational achievements and career movement of U.S.-trained doctoral scientists and engineers in the United States and abroad. The SDR uses a fixed panel design with a sample of new doctoral graduates added to the panel in each biennial survey cycle. For example, all doctorates who were included in the 2015 SDR sample and who are less than 76 years old in 2017, will be retained in 2017 survey, and a sample of new graduates who had earned their degrees were added in 2017 survey, so one person can be observed in several years' SDR sample. I will use SDR data from 2003 to 2013.

2.2 Data Processing and Cleaning

The WebCASPAR database constructed by National Science Foundation provide statistical data resources for the SED. I will use three variables in this survey: the variables of cohort, citizenship and field of study. The cohort is the graduation year of doctorate student, which is the survey conducted year of the SED. The citizenship status contains four categories: US citizens, permanent residents, temporary residents and unknown citizenship. Since unknown citizenship just contains few samples, I drop samples in this category. I also combine permanent residents and temporary residents into non-us citizens category. The field of study contains 47 majors, I make similar majors into one category and drop the majors which are not contained in the SDR, since the SDR just have the majors in science, engineering, and health (SEH) field. The final dataset contains 19 categories of majors, which are same with the majors in the SDR.

We can see the trend of doctorate recipients use the citizenship variable in the SED. The following figure shows the trend of the number of native doctoral student and foreign student from 1966 to 2000.

Figure 1: Number of Doctorate recipients from 1966 to 2000 in US



From figure1, we can see that native doctoral students increased rapidly in 1966 to 1970. Then it fluctuated from 23000 to 28000. The number of non-native doctoral students increased rapidly before 1996. In 1966, the number of non-native doctoral students was just 2500, while the number of native students is around 15000. In 1996, the number of non-native doctoral students is 13500, while the number of native students is around 28000. Nearly half of the doctoral students are non- native in 1996. The non-native students' ratio increased a lot through 1966 to 2000, which shows the increasing immigrant supply shock in labor market.

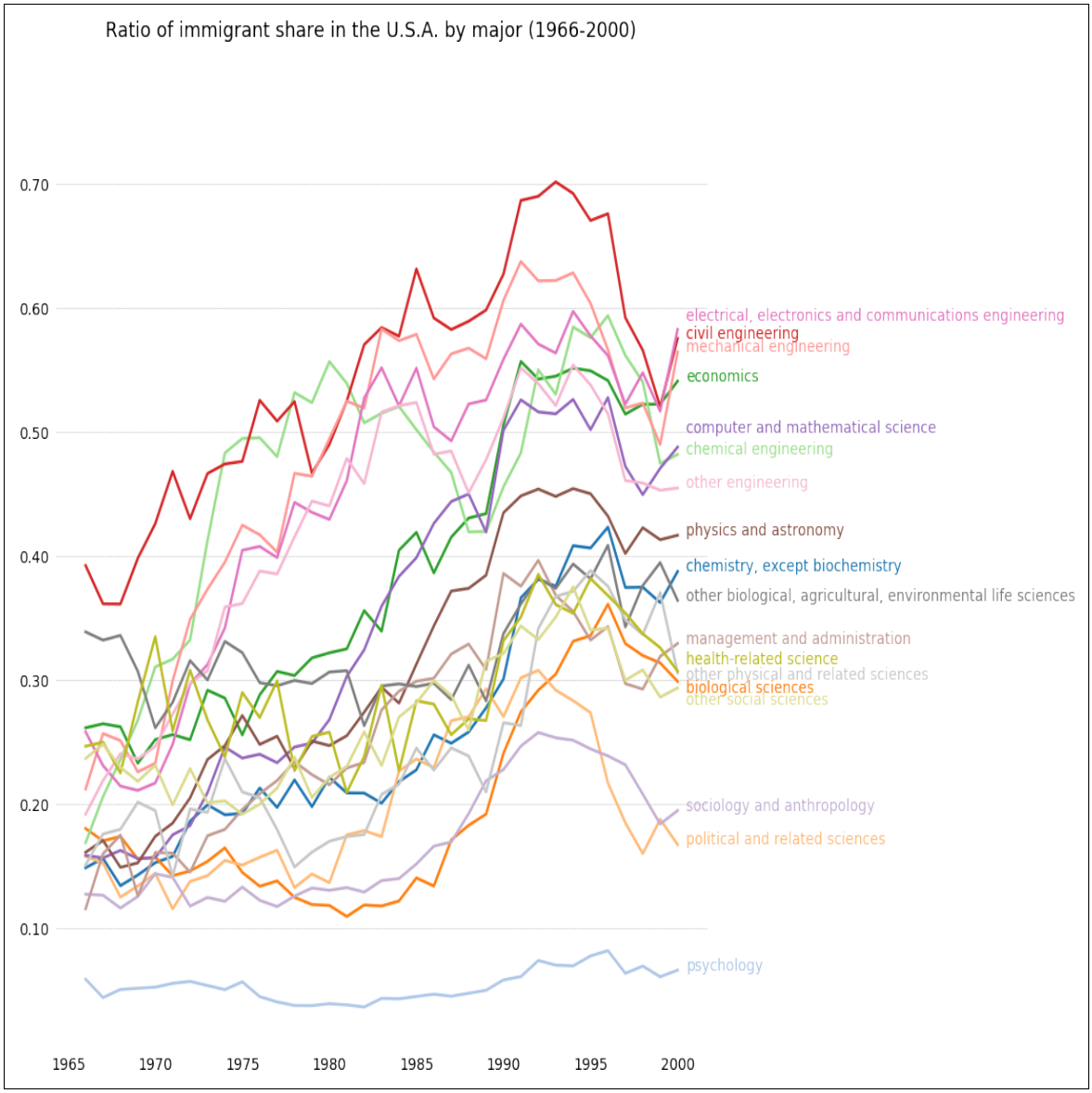
I used samples in non-us citizenship category as 'immigrant'. Then the immigrant share in each field and cohort is calculated by the following equation:

$$P_fc = N_fc/(N_fc + U_fc\ )$$

where N_fc is the number of non-US citizens, which are also the immigrants in field f and cohort c, and U_fc is the corresponding number of US citizens.

By defining the immigrant share, we can know the trend of immigrant supply shock in each field of study and each cohort through 1996 to 2000 from figure 2.

Figure 2: The ratio of immgrant share in US



From figure2, we can see that although it shows some fluctuations in these years, the trend of immigrant supply showed increase trend in each field and cohort.

The IPUMS Higher Ed is intended to facilitate access to data on the STEM labor force. I got the data from the Survey of Doctorate Recipients (SDR) from 1993 to 2003 in this database. I will use five variables in this survey: the variables of person id, annual earning, field of study, graduation year(cohort) and survey conducted year. Since the sample size in the field of 'management and administration' and 'other non-science and engineering' are too small, I dropped this two fields from dataset.

Finally, I linked the SED data and the SER data by field of study and graduation year (cohort). Since the graduation year is a large span, which is from 1966 to 2000, I divided the data into six five-year cohorts.

The summary statistics for interested variable earnings and immigrant share shows in table1.

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

3 Model and Methodology

3.1 The immigration model

The general regression model for immigration impact for labor market is as follows:

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

where $W_{ifc}(t)$ is the annual earnings for individual i who graduated in cohort c, majored in field f and participated SER in year t. $x_{ifc}(t)$ is the fixed effects of experience, which is defined as the year in the labor market for individual i. It is calculated by the year between individual i's graduation year and observed year in SER. I treat experience variable as fixed effect rather than continuous variable. Because that the intervals of graduation year is from 1966 to 2000 and the interval of SER conducted year is from 2003 to 2013, the largest experience is 47 years and the smallest experience year is 3 years. Since the interval of experience is just from 3 to 47 while the sample size is 92000, experience is better to be treated as fixed effects rather than continuous variable. d_f , y_c and π_t are fixed effects for field, cohort and observed year. This model can be added in interaction terms between field fixed effects and SER year fixed effects. This model can not be added in other variables because other variables are not significantly influence the dependent variable or perfectly colinear with the independent variables.

θ is the parameter of immigrant shock, but it is hard to interpret directly. Borjas (2009) converted it to elasticity to interpret the impact of foreign student on high skilled labor market by the following way.

$$\frac{\partial\, log\, W_{fc}}{\partial\, P_{fc}} = \theta$$

$$n_{fc} = \frac{N_{fc}}{U_{fc}} = \frac{P_{fc}}{1 - P_{fc}}$$

where n_{fc} can be interpreted as the percentage increase of labor supply con-

tributed by immigrant. Then we can get the elasticity of wage and labor supply.

$$\frac{\partial\ log\ W_{fc}}{\partial\ n_{fc}} = \theta(1 - P_{fc})^2$$

However, since one individual can appear several times in the database in different SER conducted year and p_{fc} 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 (2009) handled this model into a two stage OLS regression to make the error terms more correct:

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

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

where V_{ifc} is the individual effect in field f and cohort c and $\widehat{V_{fc}}$ is the average individual fixed effect for all individuals in field f and cohort c.

This two stage OLS regression can adjust the standard errors for the individual's correlation and the clustering of immigrant share in same cohort and field. I will use this two stage OLS regression model as base model for my research.

3.2 The bootstrap model

Full sample estimation may have overfitting problems. When maximizing the accuracy, the model may pick incorporate noise from the data. Bootstrap method is a good way to detect the overfitting problems in this two stage OLS regression and correct the overfitting problems by randomly splitting the whole dataset into training set and testing set. The randomness and repetition can help leverage the risk of picking noise from the dataset and thus correct overfitting problems.

I will measure the accuracy of the model by Mean Squared Error (MSE):

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

Firstly, I will use Borjas' model and method to do this two stage OLS regression. He used 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.

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

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: $\hat{V}_{fc}$, $\hat{x}_{ifc}(t)$, $\hat{\pi}_t$,$\hat{\theta}$, $\hat{d}_f$ and $\hat{y}_c$, and use these estimation parameters into the testing set to get the prediction of log wage. Then use the estimated log wage and the real log wage to calculate the MSE^{BS}. Then I repeat this procedure in testing set but using the estimation parameters got from full sample estimation. Particularly, in the fitting procedure in testing set, I will use the estimated $\hat{V}_{fc}$ instead of $\hat{V}_{ifc}$ in the first stage.

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

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.

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

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

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

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

4 Analysis and Results

4.1 Full Sample Estimation

Firstly, I use full sample to do the estimation and get each parameter. The result showed in table2.

Table 2: Full sample estimation results

	Annual earnings	
	(1)	(2)
Immigrant share	0.2296	0.2267
	(0.0397)	(0.0397)
controls		
(field * period) interactions	No	Yes

Full sample estimation shows that the estimated parameter is approximately 0.23. While P_{fc} is approximately 0.3 at year 2000, we can calculate the elasticity by equation that when the number of immigrants increase by 10%, the wage of native student will decrease approximately by 1.13%.

The result shows that the influx of foreign students has a significant negative impact on native students' earnings. But this estimation may have overfitting problems and we need to use bootstrap method to further study the impact of foreign student on high skilled labor market.

4.2 Bootstrap Estimation

I then randomly split the whole dataset into training set and testing set. Get estimation parameters using training set and then calculate MSE using testing set. And repeated this procedure by 10 times, 20 times and 50 times. The parameters are calculated by the following way. I don't include the interaction term of field and period at this time, since it doesn't have a significant influence on the results.

$$\widehat{\beta} \;=\; \frac{1}{N}\sum_{i\,=\,1}^{M}\hat{\beta}^{trainnig}$$

The results showed in table3.

Table 3: Bootstrap estimation results

	Full Sample	10 times bootstrap	20 times bootstrap	50 times bootstrap
Immigrant share	0.2296	0.2275	0.2286	0.2243
	(0.0397)	(0.0546)	(0.0543)	(0.0546)

We can see that θ is a little bit different by using different times of bootstrap. In 50 times bootstrap, θ is 0.2243, which is 2% smaller than full sample estimation parameter. Then 10% increase of immigrant will decrease native's earnings by approximately 1.099%. Does the coefficient 0.2243 have better prediction power and become better and more accurate than 0.2296? We also need to see the mean squared error of each method.

4.3 Compare the Mean Square Error

Using bootstrap estimation parameters and full sample estimation parameters in each testing set and calculate the mean squared error of these two kinds of estimations. The results showed in figure 3, 4, 5 and table 4.

Table 4: MSE using different estimation methods

	Full Sample	10 times bootstrap	20 times bootstrap	50 times bootstrap
$\bar{MSE}^{FS}$	0.4309	0.4187	0.4190	0.4186
$\bar{MSE}^{BS}$		0.3780	0.3775	0.3772

Figure 3: The MSE of 10 times bootstrap

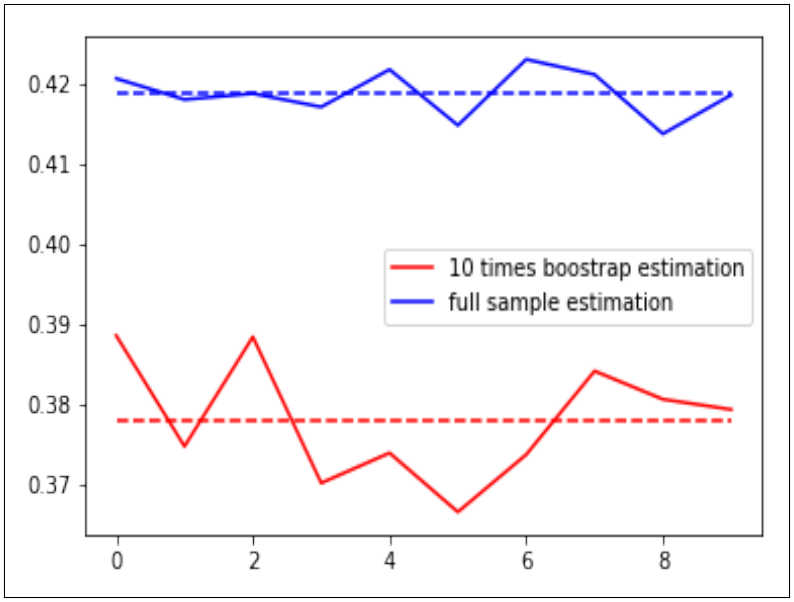


Figure 4: The MSE of 20 times bootstrap

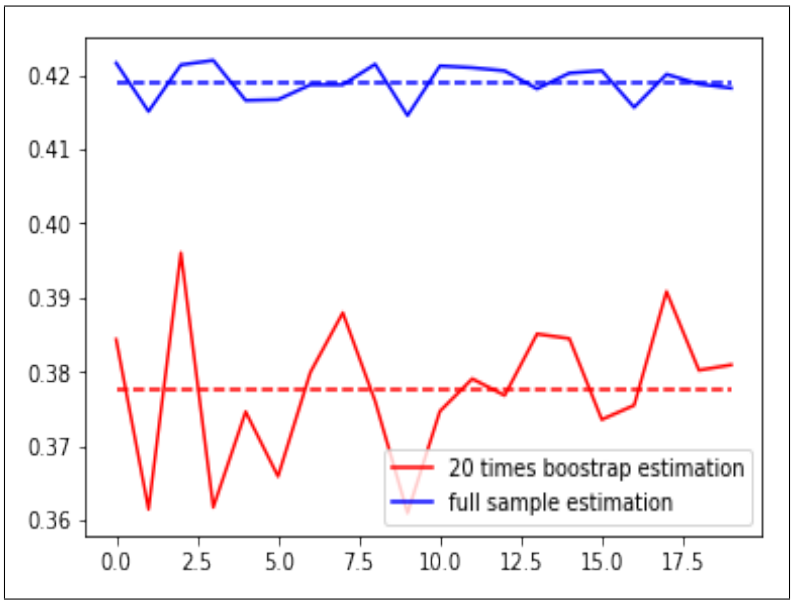
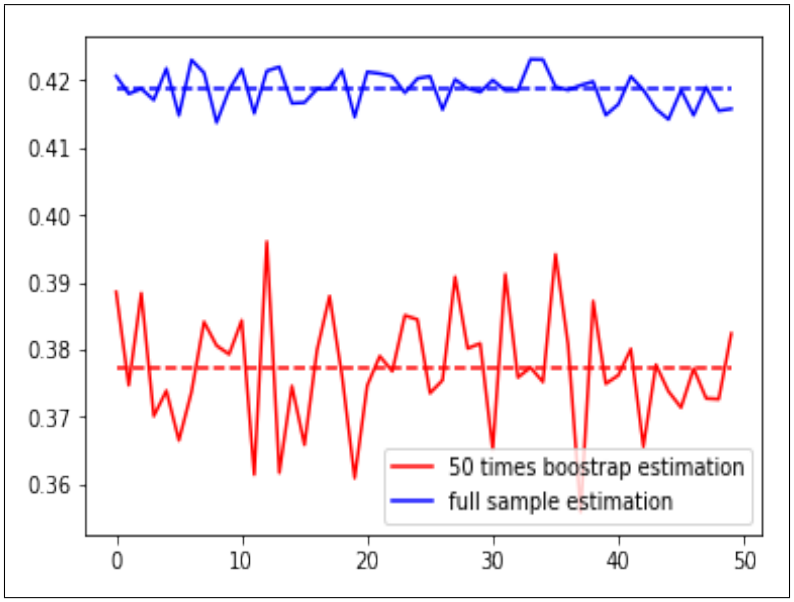


Figure 5: The MSE of 50 times bootstrap



From the table and figures, we can see that bootstrap estimation has smaller mean squared error than full sample estimation. Increasing bootstrap times doesn't influence mean squared error so much. So we can use the parameters get from 50 times bootstrap estimation. θ is 0.2243 in 50 times bootstrap.

5 Conclusion

This paper aims to use machine learning methods to correct overfitting problems in regressions. Though many researchers like to use full sample dataset to do regressions, we find that a resampling method called bootstrap can make the estimation performs better. By randomly splitting the dataset into training and testing sets, we can use the sample dataset to better mimic the full population.

This paper chooses a two stage OLS immigration model and use full sample estimation as benchmark to find whether bootstrap estimation performs better. The result shows that the mean squared error of bootstrap estimation is approximately 10% smaller than full sample estimation, which means the parameters of this two stage OLS immigration model got from bootstrap method have better prediction power than the parameters got from full sample dataset estimation. So we can draw

the conclusion that the impact of foreign students on high skilled labor market is that a 10% increase of immigrant will decrease native's earnings by approximately 1.099%.

By using bootstrap method, we successfully decrease the regression's mean squared error and find more accurate parameters with better prediction power.

6 Limitation and Future Direction

Our model will have better prediction power when using bootstrap resampling method, because the mean squared error of regression gets smaller. While the MSE changed approximately 10%, the coefficient we are interested in (θ) just changed 2%. The change of θ is not the main reason for the better prediction power. However, since other parameters are coefficient of fixed effects, it's difficult to measure which parameter caused the decrease of mean squared error.

In the future, we can expand our methodology into other regression example to find whether bootstrap method can give us more accurate parameters and generalize our conclusion. Besides, we can also explore other machine learning methods to correct the overfitting problems.

References

Borjas, George J., "The labor-market impact of high-skill immigration," *American Economic Review*, 2005, *95* (2), 56–60.

_ , "Immigration in high-skill labor markets: The impact of foreign students on the earnings of doctorates," in "Science and engineering careers in the United States: An analysis of markets and employment," University of Chicago Press, 2009, pp. 131–161.

Hawkins, Douglas M., "The problem of overfitting," *Journal of chemical information and computer sciences*, 2004, *44* (1), 1–12.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning*, Springer, 2013.

Silver, Nate, *The signal and the noise: why so many predictions fail–but some don't*, Penguin, 2012.

Steyerberg, E. W., "Overfitting and optimism in prediction models," in "Clinical Prediction Models," Springer, 2009, pp. 83–100.