Literature review:

Correct overfitting in IV regression: An immigration example

Haowen Shang

My work is based on the effort of Borjas on exploration of high skilled labor markets (Borjas 2005 and Borjas 2006). Borjas explored the impact of immigration on high skilled labor markets, specifically, the impact of immigration on doctoral labor markets. Based on the classical supply-demand theory, the influx of immigrants would increase labor supply. Immigrants would compete with native workers and thus lower the wage rates of native workers. However, some empirical studies showed that the impact of immigration on the labor market outcomes of natives is small (Friedberg and Hunt, 1995). Borjas addressed that this contradiction between theory and empirical results was resulted from the uneven distribution of immigrants across local labor markets and flows of internal migrants or capital. He also took skill group and experience into consideration. Therefore, he conducted his research in the national level and focused on high skilled labor markets. He used panel data of doctorates recipients and instrument variable regression to estimate the impact of supply shock by foreign doctoral students on annual earnings of native doctoral students. The empirical results 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." (Borjas, 2006)

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 I want to contribute to this research is to correct the overfitting problem in this instrument variable regression and find better estimation parameters.

Researchers always want to find the model that best describes the data. They may use the sample dataset to find the model with best accuracy, but this model may not perform well when used into other out of sample dataset. This is the problem of overfitting. Sliver pointed out that the signal is the true underlying pattern that you wish to learn from the data while the noise is the irrelevant information or randomness in a dataset (Sliver, 2012). Researchers want to estimate based on the 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 dataset.

Hawkins pointed out that "overfitting is the use of models or procedures that include more terms than are necessary or use more complicated approaches than are necessary" (Hawkins, 2004). If the model overfit the data, other researcher's need to reuse the modeler's software and calibration data to get similar results. A fundamental requirement of science is that one researcher's results can be generalized by another researcher in another location (Hawkins, 2004), so we need to measure overfitting and try to prevent overfitting problems. In Hawkins' paper, he also gave us an example of how to measure the fit of a linear regression model. By splitting the full sample dataset into two similar dataset--odd number dataset and even number dataset, he compared the 'resubstitution' mean square error and 'holdout' mean square error of the model. Resubstitution means using one dataset to get the estimation parameter and reapply the fit to the same dataset to get its mean square error, while holdout means using one dataset to get the estimation parameter and using the other dataset to fit the model and calculate its mean square error. If the resubstitution mean squared error is much smaller than the holdout mean square error, it is one warning sign of overfitting (Hawkins, 2004).

Hawkins gave us a simple way to measure the overfitting. This method also shows the thought of resampling--generate new sample from the dataset we have. A resampling method named bootstrap is useful for correcting overfitting problems. In the resampling chapter of 'An Introduction to Statistical Learning', the author 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 (James, Witten, Hastie and Tibshirani, 2013). 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 (James et al, 2013).

Steyerberg illustrated that bootstrap resampling was a central technique to correct overfitting and quantify optimism in model performance (Steyerberg, 2009). Optimism measures the difference between true performance of the model and apparent performance of the model, where true performance refers to model performance using the whole potential population, and apparent performance refers to the estimated model performance in the training sample. By using bootstrap method, we can randomly draw samples with replacement from the original sample. By introducing a random element of getting the sample, it can mimic the process of sampling from the whole underlying population (Steyerberg, 2009). We can use bootstrap method to get a corrected optimism estimation: the decrease between performance in the bootstrap sample and performance in the original sample. This estimation of optimism is subsequently subtracted from the original estimate to obtain an optimism-corrected performance estimate (Steyerberg, 2009). The equation of optimism-corrected performance is: "Optimism corrected performance = Apparent performance in sample – Optimism", where optimism equals to bootstrap performance minus test performance.

There are many literatures of using bootstrap method to optimize estimation. Freedman used bootstrap method into two stage least square estimation and theoretically showed that bootstrap gives the right answers with large samples even in the presence of heteroscedastic errors, which outperforms conventional asymptotics (Freedman, 1984). Fitzgerald, Azad and Ryan compared bootstrap method with standard genetic programming approach on four binary classification problems and found that the bootstrap method not only generalises significantly better than standard GP while the training performance improves, but also demonstrates a strong side effect of containing the tree sizes (Fitzgerald, Azad and Ryan, 2013). Faber and Rajko reviewed the problem of overfitting in multivariate calibration and the conventional validation-based approach to prevent it, such as cross-validation and bootstrap. They also proposed an alternative randomization test that enables one to assess the statistical significance of each component that enters the model and demonstrated the alternative approach was much more objective than traditional validation-based approach because it did not require the use of 'soft' decision rules (Faber and Rajko, 2007). Cameron, Gelbach and Miller investigated whether bootstrapping to obtain asymptotic refinement leads to improved inference for OLS estimation with cluster-robust standard errors when there are few clusters and found that rejection rates of 10% using standard methods can be reduced to the nominal size of 5% using cluster bootstrap-t procedures (Cameron, Gelbach and Miller, 2008).

I will use bootstrap method to redo the regression model created by Borjas, get new estimation parameters and find whether the predict power would be improved by using this method.

References:

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

Borjas, George J. "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*, pp. 131-161. University of Chicago Press, 2009.

Friedberg, Rachel M., and Jennifer Hunt. "The impact of immigrants on host country wages, employment and growth." *Journal of Economic perspectives* 9, no. 2 (1995): 23-44.

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

Hawkins, Douglas M. "The problem of overfitting." *Journal of chemical information and computer sciences* 44, no. 1 (2004): 1-12.

Steyerberg, E. W. "Overfitting and optimism in prediction models." In *Clinical Prediction Models*, pp. 83-100. Springer, New York, NY, 2009.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

Freedman, David. "On bootstrapping two-stage least-squares estimates in stationary linear models." *The Annals of Statistics* 12, no. 3 (1984): 827-842.

Fitzgerald, Jeannie, R. Azad, and Conor Ryan. "A bootstrapping approach to reduce over-fitting in genetic programming." In *Proceedings of the 15th annual conference companion on Genetic and evolutionary computation*, pp. 1113-1120. ACM, 2013.

Faber, N. M., and R. Rajko. "How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative." *Analytica Chimica Acta* 595, no. 1-2 (2007): 98-106.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. "Bootstrap-based improvements for inference with clustered errors." *The Review of Economics and Statistics* 90, no. 3 (2008): 414-427.