ECON 293
Class project

Luis Armona, Jack Blundell and Karthik Rajkumar

June 6, 2017

1 Introduction

In this project, we revisit two classic papers that make use of instrumental variables (IV) to derive causal estimates of quantities of interest. We look at modern machine learning methods to understand what they can add to a traditional two stage least squares analysis. The methods we look at include the lasso, regression trees, random forests, and deep IV networks. We then compare them with an ordinary least squares (OLS) approach along with other linear least squares methods using IVs.

2 Background and data

The question of causal inference in central to the study of applied economics. It is also difficult to establish because the effect of one variable on another may be confounded by several observed or unobserved factors, particularly in social science settings where one is unlikely to find experiments to isolate the particular relationship of interest. We look at two papers that tackle causal questions and provide answers using natural experiments and instrumental variables.

Angrist and Kreuger (1991) is a landmark study in providing causal evidence on the effect of schooling on individual earnings. They observe that compulsory schooling laws are usually based on staying in school up to a certain age, rather than a certain number of years. This has the effect that students born earlier in a year tend to join students born later in the year in the same school cohort, but the older students can leave school earlier because of the age-based nature of the law. Assuming that a student's quarter of birth is generally exogenous and is not correlated with ability or taste, we have exogenous variation in the number of years of schooling a student earns and we can measure its effect on their earnings. In other words, we have data on earnings and years of schooling and we can instrument this association with the quarter of birth of the student to induce quasi experimental variation in the latter, which is what we need.

Acemoglu et al. (2001) studies the effect of institutions on a nation's prosperity. They define institutions to include, among other things, the protection of property rights and the rule of law. In this context, a raw comparison of countries with rich democratic traditions and those with more "extractive" institutions is not very convincing because the establishment of the said institutions in those countries was not exogenous and depended on various sociopolitical factors specific to the history of each country. To overcome this, Acemoglu et al. instrument this effect with the mortality of European settlers during the colonial conquest of the country. The logic is that when Europeans lived longer in their destination lands, they set up strong institutions because they intended to actually live there, as opposed to places with lower mortality for them, where they attempted to extract as many resources as they could to send back home. Assuming this is valid and that institutions evolve at a snail pace once formalized, we have succeeded in finding quasiexperimental variation in institutions in countries and can then examine their effect on the incomes of people there today.

3 Methods

The workhorse tool for applied economists today is the linear regression. Under weak assumptions, OLS regression estimates have the properties of being unbiased, consistent and asymptotically normal. They approximate the best linear predictor function of the sampling distribution of the data and when a variable exhibits random variation that is unrelated to the sampling directly, the coefficient of the outcome on this variable takes the interpretation of the causal effect of changes in it upon the outcome, subject to a linear approximation of the confounding effects of all other variables present. When such variation in the variable of interest does not exist or is hard to show, one may resort to IVs, which induce variation in it and as such do not affect the outcome except through their relation to the endogenous regressor. Together, these regression models span a wide range of applications in traditional datasets.

However, with the advent of big data, there is a need for other statistical techniques as regression models are insufficient in high dimensional settings. When one has a large number of covariates, they may provide useful predictive information for the first stage of a 2SLS estimation, but this may be impossible to actually compute with regressions due to the failure to satisfy the rank condition. In these cases, regularized regression may help by introducing some bias by shrinking coefficients but also facilitating—and improving—predictive power. Further, when ones wishes to account for nonlinearities in the data, a tree structure may be better places to take nonlinear variable interactions into account and forests could help lower variance in these highly unstable models. Deep IV, in the meantime, is a modernization of the 2SLS process, where one generalizes the linear assumptions of IV to a broader class of functions spanned by a neural network model.

3.1 Lasso

The lasso started as a predictive tool in high dimensional settings to avoid noise when including many covariates by zeroing out some of them and shrinking the rest. Since then it has evolved to be used for regularization and model selection, where one runs the lasso to ask "pick out" the best covariates to put in the model. This puts one in the post-selective inference domain, because now the data generating process (DGP) model is no longer considered fixed and pre-specified, but changes with the data. In this case, work by Chernozhukov et al. shows that, under certain sparsity assumptions on the coefficients of a linear model, post-selection inference is valid in generating average treatment effects.

3.2 Regression trees

While linear regressions model the predictive function as a linear gradient on the feature space, regressive trees allow for more localization by "boxing" subspaces of the feature space and using a local predictive method (typically a version of local or nearest neighbor averaging), The obvious advantage is that one is less influenced by outliers and can better target nonlinearities in the data. But how does this help with a causal model? Note that the first stage in a 2SLS procedure is a prediction method, where ones wishes to explain as much of the variation in the endogenous regression through the instrument. There is no general reason to favor a linear specification in this leg, and one might be better able to use an instrument using a machine learning algorithm like trees.

3.3 Random forests

Given how flexible trees are, one must worry about the variance they bring to the estimates. One way to lower the variance, then, is to use the bootstrap or other resampling techniques to obtain several tree estimates and average over them. If each of these estimates were iid in some sense, then average preserves the bias and consistency properties while also improving on aggregate variance. Note, however, that while this might solve the variance problem, we still have to live with the fact that trees and forests are poor at interpolating the function where data is lacking, because unlike the OLS or other high bias tools, they are less able to use data from “far away” to make judgements on areas with scant data.

3.4 Deep IV networks

Deep IV networks form a framework to generalize the 2SLS process itself. The first stage is replaced by a nonparametric algorithm that computes the conditional density of the endogenous variable given the controls and the instrument. Using this, one uses a second machine learning algorithm to approximate the best predictor of the outcome using the above fitted conditional distribution. The second stage uses neural networks, also referred to as “deep learning,” hence the name.

4 Results

5 Discussion

References

[1] Angrist, Joshua D., and Alan B. Keueger. "Does compulsory school attendance affect schooling and earnings?." The Quarterly Journal of Economics 106.4 (1991): 979-1014.

[2] Acemoglu, D., Johnson, S., & Robinson, J. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. The American Economic Review, 91(5), 1369-1401. Retrieved from http://www.jstor.org/stable/2677930