ECON 293

Class project

Luis Armona, Jack Blundell and Karthik Rajkumar

June 7, 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.

The outline of the document is as follows: we first present background information on the two studies and the research questions. In section 3, we go through all the methods we apply to the data in hand. Section 4 shows the results we get from these methods and section 5 concludes.

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. Of course, the variation only applies to the subpopulation on whom the law binds, and not necessarily anyone younger or those intending to go to college and it would be difficult to generalize to the whole population without extrapolation.

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, Chernozhukov et al. (2015) 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, regression 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 and were formulated first by Hartford et al. (2016). 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 deep neural networks, also referred to as "deep learning," hence the name. It should be mentioned though that we only use one hidden layer in our neural network in the second stage due to computational constraints. Increases in the number of hidden layers improves fit and lowers MSE, but they also lead to volatility in the MSE estimates, particularly when the node size is taken beyond the number of features in the data set. There is thus reason to avoid doing this, especially in the "small p" setting of Angrist and Krueger. Further, our implementation of Deep IV is frequentist in its counterfactual inference and we have not resorted to the Bayesian techniques also described in the paper.

4 Results

We present the point estimates and standard errrors for the causal effects of interest computed using the various methods. (We present results on Deep IV in a separate section to be able to go through all its peculiar details.) OLS is ordinary least squares regression, IV is a 2 stage least squares regression with an instrument, and post-lasso is the post-selective inference technique from Chernozhukov, Hansen and Spindler (2015). The regression tree refers to using regression trees to partial out the effect of the other covariates (X) on the outcome, instrument and endogenous variable (Y, Z, W respectively in our notation) and running a 2SLS regression on their residuals. (Here, the outcome variable is the weekly wage a student earns after

graduation, the endogenous regressor is the years of schooling and we use quarter of birth interacted with year of birth as instruments.) The random forest averages many trees built in the above specification.

Method	Estimate	S.E.
OLS 1	0.0675	0.000337
OLS 2	0.0702	0.000336
IV 1	0.0643	0.028719
IV 2	0.0685	0.028957
Post-lasso 1	0.099	0.0201
Post-lasso 2	0.104	0.02

Table 1: Casual effects of returns from schooling computed on the Angrist and Krueger (1991) data. We have two sets of controls, one containing basic covariates and the other including all first order interactions between then. These are labelled 1 and 2 above respectively.

We see that adding the controls doesn't alter the results by very much. Further, the OLS and the IV results are very similar in this sample, which suggests there is little bias in the OLS to begin with. However, we note that the post lasso estimates are significantly bigger than the rest, which is odd especially because we expect lasso to shrink coefficients. This could have to do with the fact that we use many, many instruments are they are all only weakly correlated with the endogenous variable. This has the effect of using lots of noise as predictor variables. The other effect is that lasso selectively drops some of these variables, which could have been correlated with the outcome, but in their absence, the coefficient on the endogenous variable soaks up their presence, raising the magnitude of the effect. It may be noted, though, that the IV and the lasso coefficients are statistically indistiniguishable.

Method	Estimate	S.E.
OLS	0.308369	0.05821423
IV	1.210186	1.650509
Post-lasso	1.185839	0.5382147
Regression tree	0.3108931	0.1887891
Random forest	1.181141	0.6107168

Table 2: Casual effects of institutions on contemporary GDP per capita in the Acemoglu et al. (2001) data.

Our machine learning tools are more directly relevant in the AJR (2001) paper because there is a relatively large number of features (86, compared to the number of observations, which is 64). However, this does not also change the fact that our number of observations are low to begin with, and so we are underpowered in a lot of these analyses. One benefit that is evident above is that machine learning is better at handling "high dimensional" data: look at the difference in standard errors for the IV vs post-lasso.

4.1 Deep IV

4.1.1 Angrist and Krueger data

Let us look at the Angrist and Krueger study first. The way deep IV works is in two stages. In the first, we use a neural network to fit the distribution of the endogenous regressor using the instrument and the controls as a sum of independent normals. With this fitted distribution of the policy variable (aka W), we devise another neural net to predict the outcome variable. Given how much data we have, we are able to replicate the distribution of W pretty well.

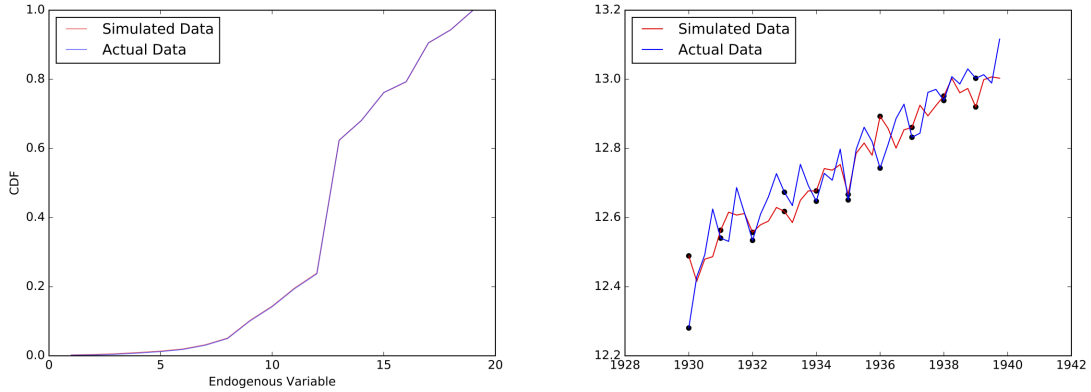

Figure 1: Left: Deep IV is able to faithfully reproduce the distribution of W. Note the perfect overlap between the distribution functions. Right: We plot the years of schooling (W) against the instrument for the raw data as well as from many simulations using the fitted Deep IV distributions. The black dots are the first quarter of each year. Deep IV does not get the seasonality perfectly but the trend is captured well as well as the general dip in first quarters.

The model parameters, i.e. the number of nodes in the neural net layers are tuned using K-fold cross validation. We show the MSE criterion for the second stage:

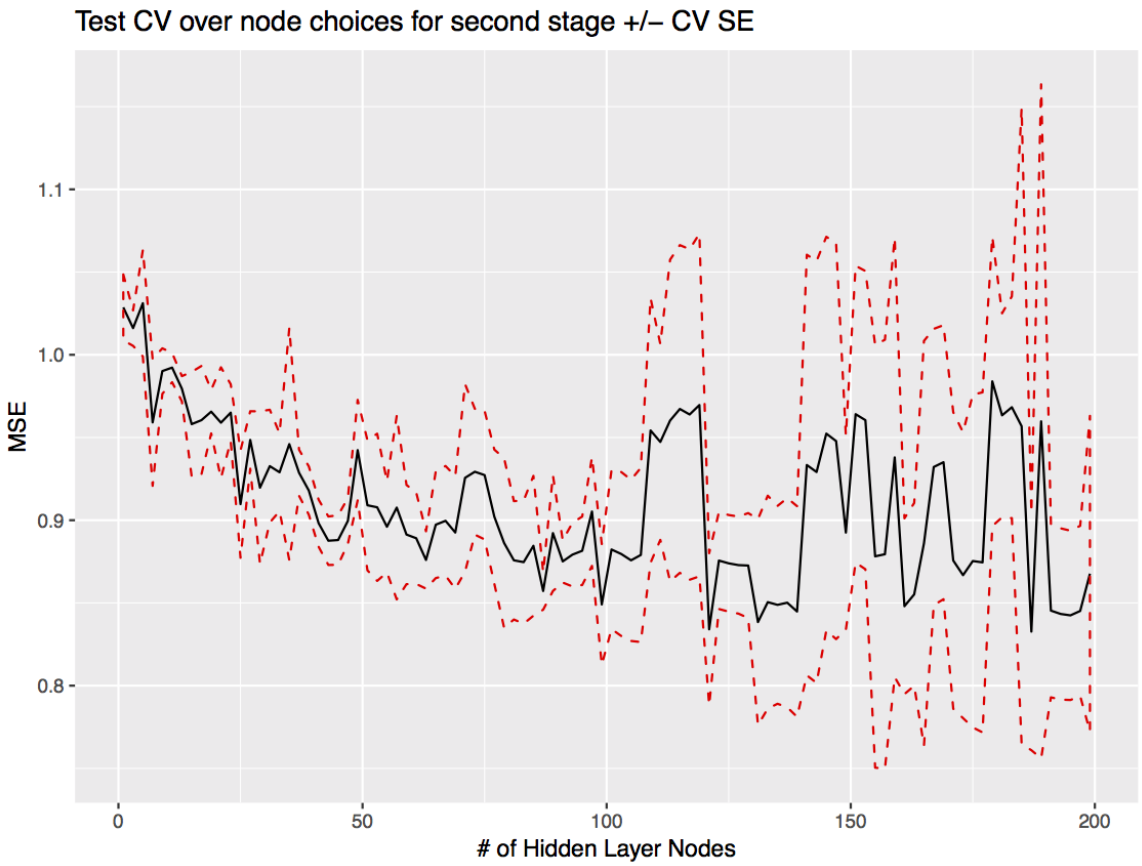

Figure 2: We pick the number of nodes associated with the smallest MSE, so long as the number of nodes is lower than the number of covariates in the dataset, 69. We decide to go with 43 nodes.

Next, we present predictions of counterfactuals using our fitted models. What is the causal effect of getting a certain number of years of education on income?

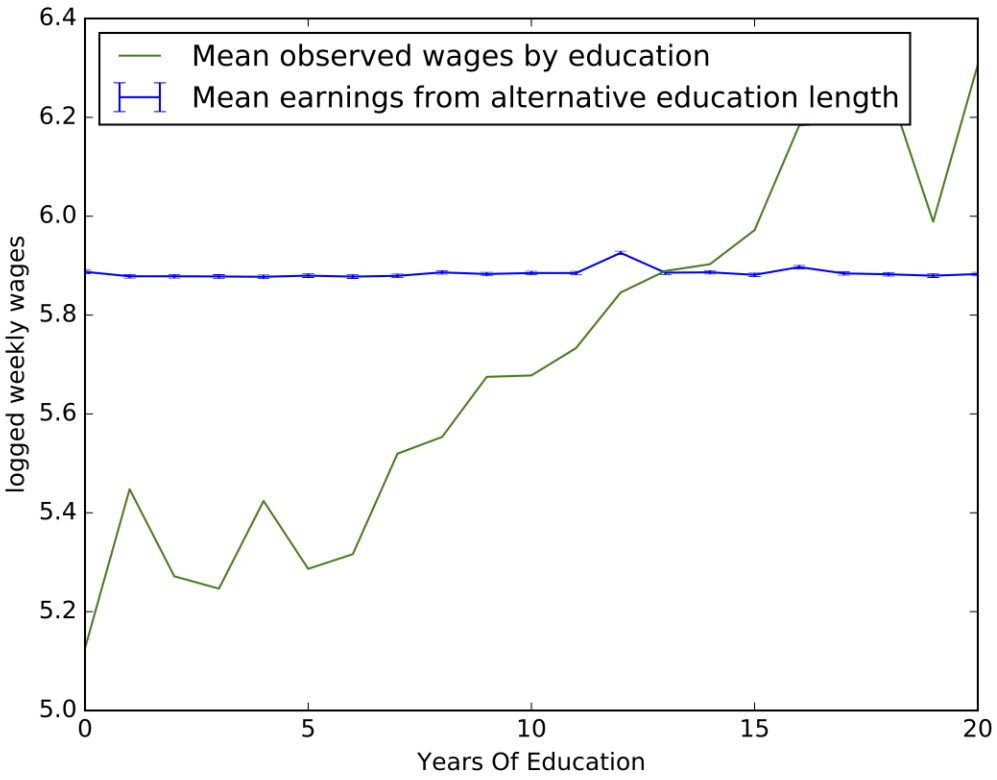

Figure 3: The green line is the raw wages observed in each cross section of the Census for people with (endogenously obtained) years of education. Note that is much steeper than the fitted blue line due to confounding biases. Further, obeserve that our "causal" blue line is actually pretty flat (at the level of the average wage in the full sample) except for one bump around 12 years of education. This is because our instrument can only really identify a local causal effect around the time of high school, which is the only time the compulsory schooling laws bind. They don't affect students in lower grades, and they certainly don't affect students interested in receiving higher education.

The beauty of the Deep IV method is that we are also able to obtain counterfactuals by varying other covariates.

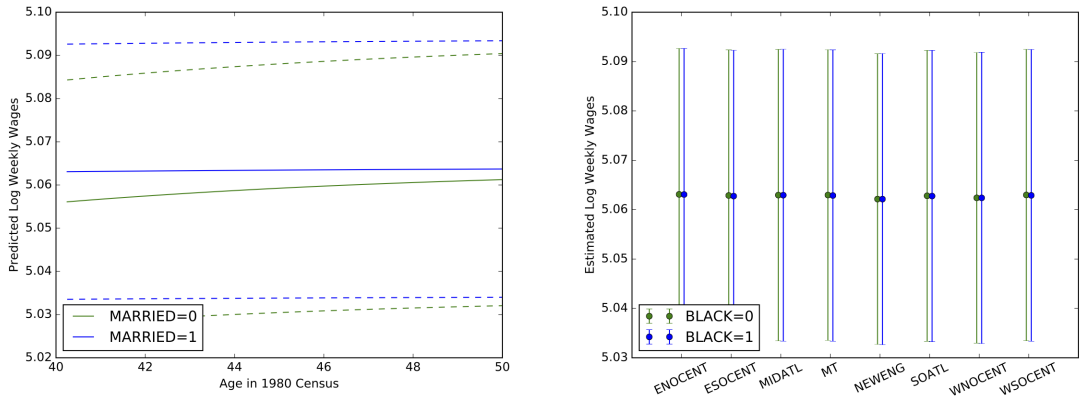

Figure 4: Counterfactuals from the Deep IV method. Left: The returns to education split by marital status. Right: Returns to education by race (specifically, whether black or not) and geographic region in the United States.

We see very little variation by demographic characteristics though, and even when we do, it is not statistically significant.

4.1.2 Acemoglu et al. data

How do we do on the Acemoglu et al. data set?

First, we plot the measures of fit in the two stages of the deep IV estimation.

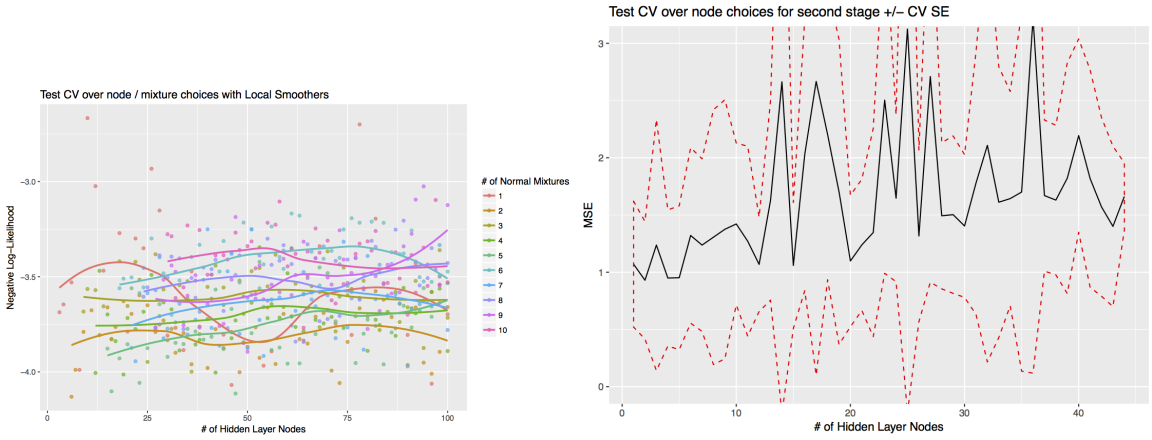

Figure 5: Measures of fit in the two stages of Deep IV. Left: The first stage uses the negative log likelihood as loss function. We pick 2 normals and 6 nodes to model the first stage distribution. Right: The MSE criterion for the structural fit in the second stage. We decide to go with 2 nodes in the hidden layer of the neural network as it minimizes this cross validation MSE.

Another way to look at the effectiveness of the fit is to compare the fitted first stage distribution with the actual data.

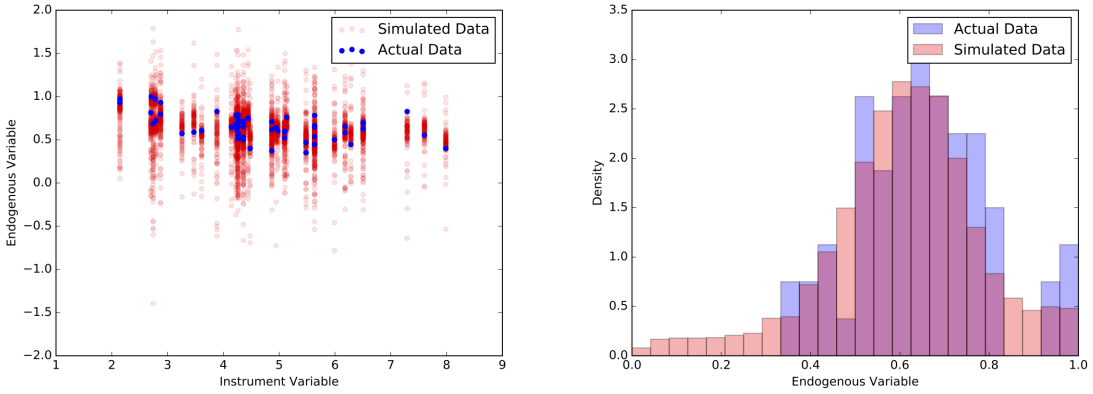

Figure 6: Left: Scatterplot of the relation between Z and W. We plot the raw data in blue and thousands of simulations from the fitted mixed Gaussian network distribution in red, noting how well the latter cover the data. Right: A direct comparison of the densities of W in the raw data versus the first stage fit. Again we note considerable overlap.

We find satisfactory performances in the fit from the two legs of the deep IV method, i.e. the first stage endogenous variable distribution fit and the second state outcome fit. Let us then look at counterfactual predictions for the question: what is the effect of institutions, as measured using an index of the risk of expropriation of personal property, on the prosperity of a nation, as proxied by the GDP per capita?

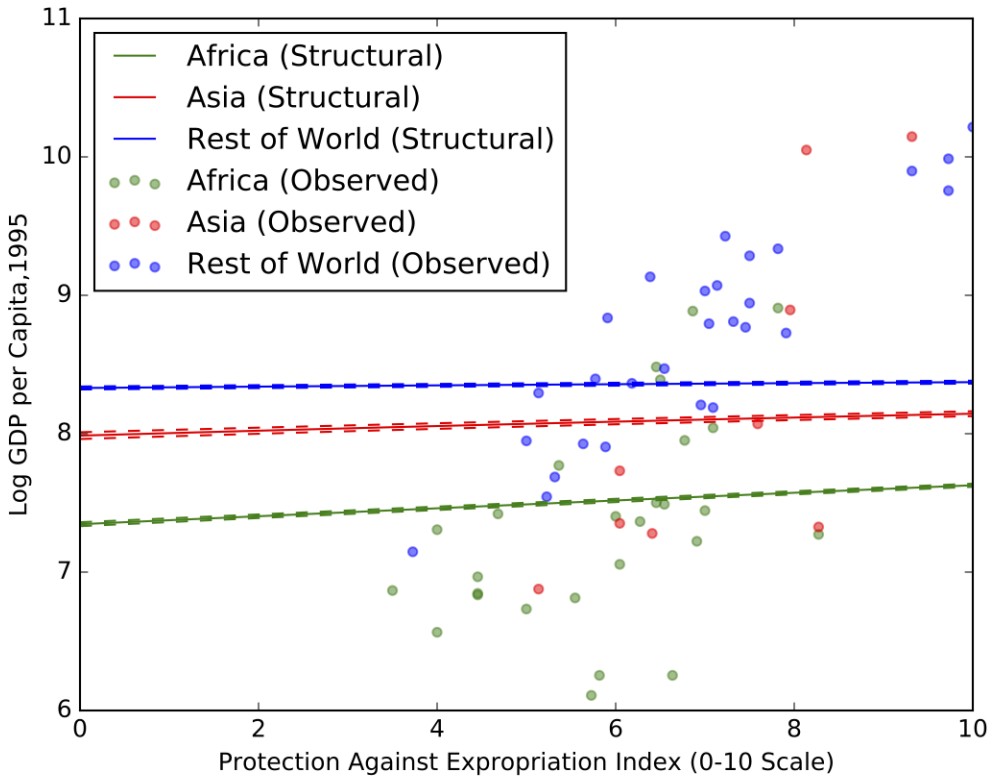

Figure 7: The scatterplot here is the raw data and the "structural" lines are causal fits from Deep IV. Just as we saw in the previous data set, the structural line is much flatter than the raw data (which does not account for confounders). Still, we note that there is a significant positive relationship between having good, institutions and prosperity. The relationship is split by geography and we note that the returns to institutions is inversely linked to the current level of prosperity of the region (Africa shows the highest) and this is in keeping with diminishing returns to investment with level of existing prosperity.

5 Discussion

We explored several new methods of performing causal inference using modern machine learning techniques. We noted the pitfalls of IV regression, including the crudeness of the linearity assumptions and the failure to deal with richer, high dimensional covariates. The lasso promises to retain the intuition with linear regression, while also cleaning up the statistical analysis to really get to the model of interest in high dimensional data. Forests are also promising because they are sensitive to local variations in the data and are able to better predict the relationships between instruments and endogenous regressors.

Deep IV promises to be a state-of-the-art method in counterfactual prediction and it is favorable in the richness and flexiblity of the counterfactual curve that can be modeled with the data, which is a giant leap from the simple one coefficient linear regression index that we traditionally use in 2SLS. At the same time, it is computationally intensive and has bells and whistles that may be difficult to interpret. Our measures of fit are encouraging but we still do not get the relationship between the outcome and the endogenous variable perfectly right. This may be because we only use one hidden layer in each of the deep learning networks.

This does not make full use of the deep learning architecture and we are limiting ourselves due to time constraints. If one had better computational resources, one could cross validate over the number of hidden layers in the neural network and get a better approximation of the underlying statistical relationships in the data. We note that we see better performance in the Acemoglu et al. data, which is the high dimensional "high p" case among the two. Deep IV performs dimensionality reduction by summarizing the information in these covariates with the number of nodes in the second stage hidden layer. The Acemoglu et al. data shows significant compression (we only use 2 nodes) but we don't see the same with the Angrist and Krueger data, which could explain why we see poor performance with the latter.

At the same time, one should remember that analysis never trumps design. The results can only ever be as good as the data and the experiment itself. As we noted with the AJR paper, we see a lot of insignificant results even with our modern methods because the data is "small" in nature along with the fact that the instrument may not be strong enough. A better statistical method cannot change that. Similarly, the instrument in the Angrist and Krueger data works only locally at the high school graduation level and thus can only really induce quasiexperimental variation in years of school in that range and nowhere else. This was clearly demonstrated in the counterfactual plot from Deep IV on the returns to schooling. One should therefore exercise prudence when replacing well understood regression techniques with modern methods and have a thorough understanding of the data and its limitations before starting statistical analysis on it.

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

[3] V. Chernozhukov, C. Hansen, M. Spindler (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. American Economic Review: Paper & Proceedings 105(5), 486--490.

[4] Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2016). Counterfactual Prediction with Deep Instrumental Variables Networks. arXiv preprint arXiv:1612.09596.