Spring 2017
MATH218 – Statistical Learning
Marcos Barrozo

Taking the `con out of Econometrics

Some thoughts from my Economics Senior Thesis

In recent years, many fields in social science have been swept by controversy related to causal inference. There is evidence that papers in Economics, Psychology (among others) resorted to p-hacking; that is, exhaustively searching for different empirical specifications until statistically significant results are found. As Ronald Coase famously said, "If you torture the data long enough, it will confess." This process leads to biasedly strong estimates. In this troublesome context, how can machine learning methods assist in better, more honest ways of doing causal inference? In this short essay, I discuss some challenges I faced in my Economics Senior Thesis. I believe my experience may lend some insights to answer that question.

In my thesis, I studied the effect of access to infrastructure on child nutrition. Specifically, I analyzed how distance to a paved road affected standardized height levels, a measure of childhood stunting. One of my hypotheses was that roads can act as a safety mechanism during a weather shock. In other words, if a bad rainfall season causes crop failures in a village, it won't be able to supply food to its local households, who will then have to import food from different regions. If this affected village is well served by roads, transport costs for these imports will be lower, which makes the impact on food prices less stark. This way, families near roads would be able to provide nourishment for their children more easily.

My analysis was concerned with finding the effect of one variable in particular – road proximity. However, as I described, the empirical process involved a variety of other variables, including household characteristics and, most importantly, rainfall. The challenge, then, was how to specify those in my regressions. My approach was just to be as transparent as possible about the process. I went through the economics literature on rainfall and nutrition and searched for how different authors defined their rainfall shock variables. Surprisingly (or not), there were many differences. So I laid out my pre-analysis plan as the following: first I'd regress rainfall on nutrition, using all the different specifications I could find. Then, with the 'strongest' result (highest p-value) for the rainfall coefficient, I'd add the road distance variables. In a similar process to what Susan Athey describes, I 'p-hacked' for the best set of covariates, and only after this did I add my independent variable of interest.

If I had a better knowledge of machine learning at the time of my pre-analysis plan, I would do things slightly differently. For the first step, of finding the optimal set of covariates, I would not test different regressions manually (and exhaustingly). Instead, I would use an algorithm to test the different specifications. Then, I would add my variable of interest.

That would be an honest way of using machine learning and 'p-hacking' in causal inference. A dishonest or ill-advised researcher could very easily just add all variables to this algorithm, finding biased standard errors. And that would take us back to Coase's argument: if you torture all the variables in your data you will find invalid results. Torture only a few of them, however, and you will have a strong set of covariates. That is the contribution machine learning methods can have in making econometrics more robust. It doesn't address all its problems (the `con), but transparency in finding strong predictors for the outcome of choice can go a long way towards a more honest research design.