Tina Chen, Geology & Computer Science, 2018
Problem Set 10, May 5, 2017

In Susan Athey's podcast with host Ross Robert, she reflected on the empirical work in econometrics, specifically citing Leamer's 1983 paper, "Let's take the con out of econometrics." This disparage of empirical data was critiqued heavily by empirical researchers back in the 80's and 90's, as he observed how "hardly anyone [took] anyone else's data analysis seriously." Athey argues that there is a fundamental difference between correlation and causality—citing current social issues like minimum wage as an example. She continues by explaining how certain methods, especially those used back in the day, had an overall fundamental dishonesty, where researchers were notorious for trying multiple different methods until something worked. And when the green light switched on, one may forget that the association is only an implication of causation, and not the true causation. This is the "identification problem"—trying to work out whether a statistical problem is caused by a certain factor when there are positive effects.

Luckily, there have been improvements in empirical work. With the advent of acquiring more data and computational power, econometrics has progressed by garnering better techniques in doing their quantitative modeling methods. Athey notes that even without "more data," the quality of methods is now drastically different, and there is a less emphasis on econometric considerations that aren't central to a causal interpretation of the main findings in question. In fact, what's driving this "credibility revolution" that Russ Robert mentions in the podcast has to do with better and more articulate research designs and methods. Recently, economists have looked into the powers of machine learning to find relationships that were not suspected in the first place.

While both machine learning and econometrics focus on finding the y-hat—the prediction value—the difference is in their underlying focus—non-causal prediction vs. causal prediction. But with their similarity in wanting to find the y-hat, why isn't machine learning used more in the field of econometrics? While economists want to be able to explain observed phenomena, machine learning techniques are data driven, and quantifying the impact of one variable one the observed phenomena is quite difficult. For example, imagine an 18-year old attending university and then earns x amount of wage. However, it is not certain what the amount of wage would be if the 18-year old had not gone to university. And by comparing one 18-year old to another, other factors would not be accounted for.

Thus, applying machine learning to econometrics can be beneficial, as explained by Athey, who has studied machine learning techniques to help isolate causal effects, so economists could draw further implications. She describes the changing the objective function, since the ground truth of the causal parameter is not observed in any test set. Similarly, when one isn't necessarily interested in the "why" to come up with an explanation, and more interested in knowing the following steps with given time to make predictions, the problem (if any), can be addressed before it happens. It is also interesting to note that their reasoning behind the research is also different—while economists are more interested in policies that comes to play, data scientists are interested in building models with the purpose of prediction. As such, utilizing machine learning to understand the covariates, to be able to classify relationships among datasets, and even identify other potential causalities can be greatly beneficial in the field of econometrics.