Ryan Rizzo

Computer Science & Economics '17

PS10

The general idea of the 'con' in 'econometrics' is the idea that it has become common practice in economics to try a ton of different variables, searching for one that agrees with a hypothesis. This leads to conclusions that are often times misleading, because they are susceptible to bias in the dataset. So many of these studies have real-world effects, but the economists are just trying to get a paper published, so they search until they find significant p-values.

Susan Athey proposes a different standard, a way to combat this trend in many empirical studies. Her idea bases around the fact that if you give a computer thousands of variables and tell it to find trends, it will find things that aren't really there, and it will find things that are really there, but there is no way to really tell. It is comparable to an economist with unlimited time searching through piles and piles of variables finding any and all correlations. But the most important part of Athey's proposal is to use the best model found by the computer on a new set of data. That way, if the model holds true without bias from the previous dataset (or one half of a dataset), then you know that the model holds merit. This, however, does not mean that there is a causal relation. This machine learning predictive model is not about causality, but about fitting a model based off of as many variables as possible in order to project future results.

The reason this predictive model cannot be used to show causation effects is because any given variable could be correlated with another variable which holds the actual causality. For example, in Leamer's "Take the Con out of Econometrics," he talks about how crop growth could be attributed to shade from trees, or to bird droppings from those trees. Either one could be the answer, or the answer could be both. But any two economists could come back with different answers. That's why machine learning algorithms can predict so well- they can sort through so many variables in a matter of seconds. But in even the largest data sets, there still could be some bias with some variables, and that is why it is important to test those models on a new set of data.

But if we combine machine learning with econometrics, we can find those variables which run counter to the causal correlations that the economist finds. Machine learning, on its own, is somewhat of a black box. Econometrics, on its own, is an economist picking and choosing what will tell a good story. But together, we can get closer to the "true" story.