Nina Sonneborn
Mathematical Science, '19

How can machine learning methods can help take the 'con' out of 'econometrics'?

Traditional econometrics is skilled at modeling small data to show causal relationships. Economists' models are interpretable, while machine learning struggles with modeling for interpretation. The strength of machine learning methods lays in Big Data and accurate prediction. As proposed by Susan Atley in her EconTalk podcast, machine learning methods can be used in econometrics as a way to understand covariates and to develop a systematic approach to model selection. The 'con' of econometrics is that with enough data and enough variables, it's easy to find some sort of test on some variable that shows a given desired outcome. This shows a "causal relationship" where one may not actually exist because it doesn't accurately control for lurking/confounding variables. By combining machine learning techniques with traditional econometrics, we create more honest models.

Machine learning methods involving sample splitting can be used to determine which variables to use in the economic model. Sample splitting partitions the full data set into two parts – one part, called the training set, is used to train the model, while the other part, called the validation set, is used to test the model. Essentially, this keeps the statistician honest, making sure the model is not too overfit to the data. Of course, sample splitting can sound unappealing when working with a data set that is already small in size. However, more complicated machine learning methods allow the modeler to incorporate sample splitting without having to use only half of the data to train the model.

This also allows for data-driven model selection. Machine learning methods can analyze a large number of covariates and let the data say what's most important. Then, these variables can be put into the final regression. When variable selection is data driven, it is less susceptible to human bias.

Machine learning can also be applied to create synthetic control groups in order to determine a single causal relationship. Machine learning techniques can be applied to find the model that best fits the untreated data (the control group). This model should not be given any causal interpretation, but will have a very good fit to the untreated data, which will make it easier to separate out the effect of whatever singular variable whose effect the economist is trying to measure. The creation of a counterfactual can be difficult, but since the objective is to interpret the effect of the treatment rather than the other variables, models that have good predictive power but are hard to interpret can be used to sort out the effect of the treatment from other variables that aren't of interest.