Ben Czekanski
Mathematics and Economics 2018

Taking the "Con" out of Econometrics with Machine-Learning

The questions that econometrics asks are very important. As discussed in the podcast, "Does X cause Y?" is the fundamental question that scientists try to answer, and econometricians are no different. Causality is generally proven using differences between the observed and the theoretical counterfactual. The outcome that is most often used to prove the strength of the conclusion is the p-value. However, p-values can be "hacked" by changing the specifications of the experiment. This usually involves selecting a different subset of the data or trying different statistical techniques until you find one that shows a relationship that isn't really there and doesn't show up under other specifications. While econometrics has gotten better about "p-hacking," this is where machine-learning techniques can be used. The idea of cross-validation is one that could be productively adopted by econometricians. Rather than testing each specification manually, and using human judgment to determine which one makes the most sense and works best, cross-validation can be used to determine the model specifications that deliver the most consistent results across multiple subsets of the data. This will introduce a sense of out of sample relevance that econometrics is missing. Generally, econometricians delve deep in to the data that they have, but really struggle to persuasively state that their findings can be found outside of that specific setting.

Another way in which econometrics can be improved by the introduction of machine learning techniques is by introducing the concept of predictor selection. Econometricians usually do the best to select the variables that they believe will be the most theoretically relevant to the outcome that is being investigated. The computer can do this task through methods like regularization. The other element of variable selection that could be incorporated is the concept of selecting covariates. A tendency of econometrics is to pick manually the variable that is theoretically the most representative of a group of correlated predictors. This is the high-dimension problem that was discussed on the podcast, where machine learning can empirically determine the best predictor or even possibly create a composite of the correlated predictors that can be used as a predictor on its own, incorporating all of the related predictors. As discussed on the podcast, while having its issues, econometrics is not fundamentally flawed; it just can be improved with the incorporation of machine-learning techniques. Many of these techniques simply automate tasks that are usually done by the researcher, and can determine factually the best option for decisions that were usually left up to the intuition of the researcher. For anyone working with data, both machine learning and econometrics have important techniques for making sense of large quantities of data.