Brenda Li, Mathematics, ‘17

The goal of an economist is to explain how the world functions. Whether it’s investigating if an increase in minimum wages boosts employment or what the true economic return of higher level education is, economists use data in order to try and create reliable models that produce informative and unbiased results. However, the reality is that many statistical methods used in econometrics as well as other research disciplines are quite problematic. One primary issue with many econometric methods is that researchers are often overly preoccupied with finding causal relationships between the variables they’re analyzing in order to confirm a hypothesis. Here, the problem lies in the fact that if you analyze enough things in a large enough variety of ways, then you can find any kind of result that spuriously confirms your hypothesis. Thus, there arises many inaccurately and statistically insignificant interpretations from biased models. However, this is where machine learning can provide a solution for this issue by offering systematic ways to create models and evaluate the validity of a hypothesis. In particular, Professor Susan Athey argues that two strategies in which machine learning can address this issue of problematic causal inferences is through automating the model-making process and using sample splitting to evaluate the error of a model.

According to Athey, many of the issues arising from overanalyzing your data for significant trends could be solved by letting a machine create models for you. By programming a computer to create your models, not only is there no unsystematic bias in the way predictor variables are utilized by the model, much larger datasets can be used due to the machine’s computing power. Furthermore, a computer is able to automate the model making process so that you can ask it to create and test tens of thousands of model in order to determine which model performs best. This way, the models generated are more accurate and have much greater predictive power. Furthermore, Athey also proposes using sample splitting as a method of making sure your model isn’t fit too closely to the training data you used to develop models. In other words, Athey suggests that researchers should separate their incoming dataset into two parts; one part goes into training the model and the other part goes into testing how well the model performs on new data it wasn’t trained on. This sample splitting method helps addresses the issue of problematic causal inferences because it provides a way for researchers to determine whether their predictive models are too biased toward the dataset that was used to create it.

While automating the model-making process and using sample splitting alleviates some of the consequences of problematic statistical modeling, these machine learning methods are not the end-all solutions and they have their own drawbacks. For one, these models are not very interpretable as it is difficult to make inferences about the relationships of the variables involved in the optimal model when the variables were selected by a computer more or less at random. Furthermore, these models may not be reliable across different circumstances as they still risk being overfit since the model relies on nothing but the dataset you feed it. Since interpretability and reliability are two of the main principles of econometrics, Athey urges researchers to aim for the middle ground between traditional economic analysis and machine learning. She advises that machine learning should be utilized to automate the model-making process to generate better models, but some parts of the systemization of the machine should still be constrained so that models are reliable and interpretable.