

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

In Dr. Susan Anthony’s hour-long interview on the Econ Talk podcast, she discusses how to predict the effects of increasing minimum wage in a tech-based city. In doing so, she exemplifies the ways in which machine learning can improve statistical modeling. Essentially, Dr. Anthony describes the ‘train’ and ‘test’ sets of data we receive in class. By creating a model without biases associated with data mining (false positives and false negatives), predicting the real effects of a variable on an outcome variable becomes higher. The example used in the interview is that of a drug given to 1,000 patients. We know 1,000 variables about each patient and we know 15 patients have extremely good outcomes. Traditional data mining would find what variables the patients had in common. Machine learning runs for loops to uncover the ‘model of best fit’ and tests that on a second set of data. Its data mining in a rigorous, replicable way! Historically, researchers could design an answer before they had asked the question. Rather than creating an answer based on covariates that the 15 patients had in common, machine learning determines what relationship between variables is most accurate across the entire dataset.

I thought the most interesting fact in this podcast was that many many many statistical studies, especially in psychology, are unable to be replicated because they were created without machine learning, they were created with data mining. This has had HUGE implications on pretty much everything we eat, say, and do. Historical institutionalism means that our shared knowledge is layered on top of itself as it is created – what would it mean if layers and layers of that data was just...wrong?