Otto Nagengast

Economics, ’17

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

In our econometrics courses, we learn that correlation is not causation. Say that we want to find out if education has an effect on a person’s income. We would begin with the “naive” regression of income on education. We find that education has a highly statistically significant, and large impact on earnings. Can we conclude that more education results in higher income? No, because there may be omitted variables. These are factors that covary with education and have an effect on income. Parental income is an example. Someone that receives more education likely has wealthier parents. But coming from a wealthier family may also give the child more lucrative job opportunities because of their family’s network. In our model, the coefficient on education would actually be picking up the impact of education *and* parental income on an individual’s income. To isolate the impact of education, we need to control for parental income by including it in our regression. We continue this exercise until we control for all possible covariates of education that also impact income. We can then interpret the coefficient on education as the causal impact of education on income.

The ‘con’ in econometrics that Athey and Roberts discuss is called p-hacking. When we test if a coefficient is statistically significant, we rely on a t-test which generates a p-value. A p-value tells you the probability that you would find a value at least that far from the mean by chance.[1] In economics, the standard threshold is a p-value of 5%. If the p-value on education’s coefficient is greater than 5%, we are forced to conclude that education’s coefficient is not different from zero; that is, education has no impact on income. A p-value of 5% means that in one out of every 20 times we fit this model to a random sample from the population, we would expect to see a value at least as large, in absolute terms, as the coefficient on education that we found. This would happen just by chance. P-hacking tries a large number of different combinations of controls in our model to find the set of controls that produces the lowest p-value for our variable of interest. Eventually, a particular version of our model will produce a p-value for education’s coefficient that is less than 5%. We could publish this model and show that we found that education has a causal effect on income. But this would be dishonest, because the statistical significance of education’s coefficient is a product of chance.

Athey proposes using p-hacking to identify the best model and then testing this model on different data. The process looks as follows: First, we would split our data in half. On one half of the data, let’s call it the training set, we use machine learning methods to find the set of controls that give us the lowest p-value for education’s coefficient. Then we take this model and apply it to the second half of the data, which we can think of as the test set. If education’s coefficient is statistically significant on the test set, then we can be confident that we adequately controlled for possible confounding variables. We may not be able to interpret the coefficients on our controls, but we do not care. All we want from our controls is to control for all the things that covary with education and that impact income.

In econometrics, we spend a lot of time thinking about possible confounding variables and how we could control for them in our model. This often amounts to guesswork. If we have data for these variables, we can let machine learning do this work for us. These methods will systemically find the best set of controls for our model, which is better than us trying to guess.

[1] This is for a two-sided hypothesis test.