Alexander Pastora ‘18
Molecular Biology and Biochemistry

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

According to the podcast, “econometrics” refers to the use of a model to determine causality via the assessment of covariates, as opposed to using a predictive model. By taking the “con” out of “econometrics” we are referring to Susan Athey’s claim that machine learning methods are being used to look for trends in data sets that are super specific to the data set being looked at, and do not really hold true when the model is applied to a new dataset. As Athey stated in the podcast, if you are looking at one outcome variable and thousands of predictor variables, you are bound to find some sort of trend that exists in your data, which you can then justify with an appropriate p-value. However, when these models are being created, the whole data set is being used, so there is no way to determine if the model itself is actually good at taking in new values. At times, these models are not even able to be reproduced by researches testing them out. This, of course, become problematic, because this type of model fitting has been historically used in various fields of study, such as Economics and Psychology. This is exactly why Athey wants us to change the way we use machine learning methods.

To begin taking the “con” out of “econometrics”, Athey claims that we must first create synthetic control groups. That is, we must first take our data set, which has covariates that are changing, and we must compare it to a similar data set or data sets where the covariates are stagnant. By doing this, we can better see how changing covariates affects the trends in the data set. Specifically, in the podcast, Athey explains by taking the city of Seattle, and determining the benefit that raising the minimum wage has had in the city. In order to do this, however, we must compare Seattle to different cities that did not have a raise in minimum wage at the same time. By doing this, we can better determine that Seattle’s raise in minimum wage actually had a significant affect. Additionally, Athey tells us that we must really listen to the data when we are attempting data mining. That is, we should use the data to both make the model, and to refine the variables in the model to improve its accuracy.

According to Athey, data mining can both be good and bad. In order to use data mining correctly, one must split the data set into pieces, and use a portion of the data to fit the model. This is so that the model is not overfit to the data set at hand. By doing this, it is then possible to evaluate the model on the remaining portion of the data set. In fact, although not directly stated by Athey, it is possible to fit the model, and then use cross validation methods (via data splitting) to verify that predicting ability of the model is good. However, as mentioned by Athey, you cannot really assess the “goodness” of the model until you introduce a completely new data set. Overall, the goal of reducing the “con” in “econometrics” through machine learning methods is to increase the predicting power of the model, and to create a model that can hold up in varying circumstances.

In my area of study, it is kind of difficult to take the “con” out of “econometrics”, because we are not actually creating a model using data (this is definitely more of an Ecology or Evolution thing). However, I do think that these ideas are very relevant to Microbiology and Genetics research. Often times, you are looking at a specific gene and seeing if it is associated with the production of a certain product. One way you do this is by either knocking out the target gene, and seeing if the production of the product decreases, or transfecting the gene into different bacteria lacking that gene, and seeing if the product is formed. In this case, when you transfect a gene into new bacteria (via a plasmid), you must create a control group with an empty plasmid to determine that the plasmid has not changed anything. As in Athey’s example, if you want to introduced “new data” into the system to “test the model”, you would transfect the gene into different bacteria to see if you get similar results.