Emily Y. Miller
Molecular Biology and Biochemistry ‘17”

In Econ Talk’s interview with Stanford Professor Susan Athey on the use of machine learning in the business and tech industry, the moderator Russ Roberts asks Athey how machine learning methods can take the ‘con’ out of econometrics. In the podcast, Roberts stated that the ‘con’ in econometrics is when if you are ‘on a fishing expedition’ and running data that is exceptionally varied, the classic measures of hypothesis testing and significance analysis do not hold, and therefore most researchers will chose a model that confirms to their biases. Athey gives an example of how machine learning can be used to create artificial counterfactuals or synthetic control groups that are able to systematically evaluate the validity of assumptions.

While the portion of the podcast that focused on economics was interesting, what I found most relevant was the portion of the podcast that focused on how machine learning can fix “bad” statistics. I think the real issue that underlies the “con” in econometrics and bad statistics is that one statistics class does not prepare a person to conduct anything more than the most basic statistics, and it is easy to go beyond one’s knowledge base. As an example, I would like to take this time to point out some curricular issues in the Middlebury College Biostatistics class that I believe could be contributing to students’ ignorance of statistical processes and could lead to Middlebury College biology students ‘conning’ themselves, their future supervisors, and anyone who reads their research into believing there is significance when there is none. For instance, during my Jterm 2015 Biostatistics class, I was taught that if the data displayed homogeneity of variance a parametric statistical test (one that also assumed normality) could be used, and there was no need to check and see if the data was normal. Tests for data normality were neither mentioned nor taught in this class. While this is perhaps the most flagrant example, the class also did not teach how to do a sample size calculation so that a study has enough power to form statistical conclusions, or that there is a multiple comparisons problem when performing multiple statistical tests.

The main problem I see with undergraduate statistics education is that when one takes a single statistics course, as required by many economics, biology, or psychology majors, that person knows just enough statistics to be dangerous. Furthermore, the ease of running statistics in current programs (SAS, R, STATA, etc…) makes it easy to never understand the math behind statistics and even easier to go beyond your knowledge base without realizing it. The current statistical computing programs make it easy to trust the output of a statistical test without checking or even knowing the impact of violating the assumptions the test was based upon. This is one area where machine learning can improve statistics as the role of assumptions in machine learning is much smaller. Some students will get to take further statistical classes to remedy the gaps in my knowledge. However, many Middlebury College students will go on to research jobs in the next year where they may be expected to conduct statistical tests. If machine learning can take the ‘con’ out of econometrics, bad statistical teaching and overconfidence is what put the ‘con’ there in the first place.