Shannia Fu ‘17
Mathematics & Computer Science

In any kind of scientific statistics-based experiment, it’s important to control for confounding factors. In general, to do this kind of testing, there are control groups with which to compare test groups to see how different factors affect the results of the experiment. However, it is almost impossible to control for every possible factor or variable. That’s where machine learning comes in.

Nowadays, because there is so much data in the world, there are often times where a data set has more characteristics, or predictors, of each observation than it has total observations. In this case, one can find so-called trends in tiny subsets of the data, even if they are actually non-trends. In other words, people running an analysis on the data may find an explanation for a phenomenon when, in reality, the phenomenon might not be significant or even exist, or the explanation is too specific to the data to be able to applied to a wider set. And this is where confirmation bias can come in. The hypothesis that the experimenter held before beginning the experiment would be “found,” when not actually statistically significant, only because the experimenter was already biased towards that result. One method in the field of machine learning helps combat the problem of having too many predictors or factors in data. This method, called regularization, helps pick out variables that actually matter, by fitting regression curves on data sets over and over with a different penalized variable each time.

An example where this problem arises is in drug testing. Say, for example, a drug works really well on only a small set of people that you have a lot of data on. Then there will definitely be at least a few characteristics you can find in common between each of the successful cases. And you would conclude that the drug works really well on people who share those specific characteristics. But to prove that this hypothesis is right, you would find just a couple more people who share these characteristics, even if only slightly, with the same good outcome. But it could have just as easily been a random 15 people who had good outcomes. You don’t know, because there were too many variables.

To test your regularization model, then, split your data in half to fit the model on one half and test it on the other. After testing it on the other half, you can confirm whether your model is valid.

With this modeling method and its accompanying method of testing, the effects of both overfitting (i.e. fitting a model too specific to too few points of data) and confirmation bias are diminished.