Lecture 11 – Prediction

Prediction and Causal Inference have different but (often) complementary goals. With causal inference we are interested in the effect of some "treatment", D, on some outcome Y. We are not interested in knowing everything about Y, how to predict it, etc–we're just trying to get the effect of D on Y. In a prediction problem we are interested in being able to predict Y and are less concerned about knowing the effect of D on Y. One way to say this is that causal inference is interested in a $\hat{\beta}$ whereas prediction is interested in a $\hat{Y}$

A key issue is determining if you are asking a causal inference question or a prediction question. I have seen this called an "umbrella" (prediction) versus "rain dance" (causal inference) question. Both of these types of questions are useful depending on the context.

Admittedly, causal inference does involve prediction–essentially you are trying to predict a counterfactual for your treated group. This could be a synthetic control group, a control group in an RCT, the untreated side of the cutoff in an RD, etc. However, this differs from trying to predict which applicants will default on a student loan, which restaurants will fail an inspection, or which students will graduate from college.

Let's think through some examples of problems and if they are about causal inference or prediction.

1. If a city expands its pre-K program, will children have higher test scores in high school?
2. Which high school students are most at risk of dropping out?
3. If the state raises the minimum wage, will teen employment fall?
4. Which restaurants are most likely to fail a health inspection?
5. Does adding police patrols to a neighborhood reduce crime?
6. Which households are likely eligible for SNAP but not enrolled?

Thinking about who is asking the question and what they control is pretty useful. A school district can't change a student's family background, but it can use that information to predict who is at risk of dropping out and target counselors to them. This is a prediction exercise – the district needs an umbrella, not a rain dance.

On the other hand, a city council deciding whether to fund a pre-K expansion controls that decision directly, so it needs the causal effect of the program.

Your previous statistics and econometrics coursework covers a fair amount of prediction (even if it doesn't call it that). You can run OLS, a logit, or probit, and then take the estimated β's and predict what will happen.

It has become fashionable to call prediction "Machine Learning" rather than prediction but the basic ideas still hold. The goal of this lecture is to introduce you to the basic terminology used by people doing machine learning and show how it maps into what you have learned to this point.

Terminology/Intuition

Essentially all data science takes some source of data and uses it to **train** the model. This means estimating coefficients (basically). It then checks how good the training is in some sort of out of sample data set. Often this means that you pick some subset of the data to train the model (**training data**) and then some subset (**test data**) of the data to test the model.

This step is very useful because what does OLS or Logit (or basically any prediction algorithm) do? In the case of OLS this minimizes the sum of squared errors. This means it generates the best possible fit it can in that data set. But, some of the stuff in the training data set is just noise rather than useful in prediction. So, we don't want to **overfit** the data where we put too much weight on noise.

Why do we care about overfitting? We do not care about predictions in sample, we care about predictions out of sample. Being good at predicting in sample is easy–we already know the answers! Prediction is only useful out of sample. Out of sample could be a similar data set collected in the same year or it could be trying to predict what will happen in the future.

Ex) Predicting Causal Inference ability in our class

Some problems want you to predict something that is binary. This is called a **classification problem** (e.g., will a defendant fail to appear in court?). If instead, you're interested in predicting an outcome that is continuous, this is called a **regression problem** (e.g., how many ER visits will a Medicaid enrollee have next year?). The way you implement any specific algorithm is going to depend on whether it is a classification problem or a regression problem

When does causal inference meet machine learning?

Increasingly, machine learning is being used in causal inference. Susan Athey and Guido Imbens (and others) are working on combining insights from causal inference and prediction. In many settings, you may need to predict who is treated and then use that as an input to a causal inference method.

Ex. Top 10 Percent Rule