Prediction vs. Inference

So far, our statistics material in this course has fallen into two buckets. The first, and most straightforward, is *descriptive statistics*, that is, just describing what our data looks like—mean, median, correlation, that kind of stuff. The second is "inferential statistics," that is, use of statistics to make inferences about unknown population parameters (often on the basis of some kind of causal theory). Here's a nice Statistics by Jim explanation of the difference.

But there's a third major use for statistics, which is prediction. Fundamentally, prediction is about what's known as "out of sample data"—that is, data that you haven't seen, and figuring out what values to attribute to it.

The easiest way to think about prediction is to think of a linear regression again. If you have a simple linear regression, like, house prices on square feet, that doesn't just help you test a causal hypothesis. It helps you predict the prices of future houses, just from their square footage alone. Once you get your regression line, you can just multiply the number of feet by the coefficient you get from your regression, add the value of the intercept, and, on the basis of your data, you've got your best prediction of the value of the house.

It turns out that prediction is extremely useful. In industry, this is basically what data scientists do on a day to day basis (although they use much fancier mathematical and computational tools than simple linear regression). For example, suppose you've got a website, and you're trying to figure out how to maximize the sales of some product through it. One common thing that industry data scientists will do is called an "A/B test," which, essentially, is just an experiment where they create different versions of the website, and then see which versions lead to more sales. Then they can predict future sales, and take the highest-value version of the website.

But things get much fancier. For example, there's a discipline in machine learning (which is mostly a fancy name for predictive statistics) called natural language processing (NLP), which involves, among other things, figuring out how to read texts written in human language. Here's how it works from about 10,000 feet up.

Suppose I have a big dataset of tweets, and I work for a company that wants to be able to figure out whether people are mad about our products. I can do "sentiment analysis" on those tweets by hiring humans to *label* a large number of tweets as either positive or negative ("Windoze SUCKS!!!" ok that's negative. "I LOVE MY IPHONE!!" pretty positive). Then, we can represent tweets by the words that are in them (imagine a dataset where every possible word is a binary column, 1 if the tweet contains the word and 0 if it doesn't), and fit a logistic regression (remember how I said that's what we use instead of linear regression when we have a binary dependent variable) on that dataset with the dependent variable being a binary column that represents our labels (1 if the human said the tweet was angry, 0 otherwise). And then, if I've done my job right, the company can continually track how angry twitter users are about its

products by sucking in tweets with the company name or the name of any of the products as they get posted, predicting their probability of being angry using my logistic regression, and then calculating up the ratio of angry to not angry tweets.

Actually, real-life natural language processing is much more complicated than that. (For example, we wouldn't actually represent tweets as simple presence/absence columns of words; these days they would actually be represented as vectors of values in a high-dimensional space of, effectively, linguistic meaning as observed correlations from a different machine learning model... Google "Word2Vec" if you're curious.) But this gives you the basic idea. There are also computer vision researchers doing similar things, but with pictures instead of text—making it easy to do things like differentiate between cats and dogs on photographs over the internet, or, on the polar opposite end of sinister-ness, making it easy for governments to do facial recognition to identify people from photographs.

This stuff has many applications in the law. In terms of practice, the most advanced, in terms of commercialization and existing integration into legal workflows, is probably predictive coding—which you can represent as basically fitting a classifier (something like logistic regression with binary or multi-class outcomes rather than continuous outcomes) to a set of discovery documents to determine if they're responsive or not responsive (or privileged/not privileged), hence potentially saving vast amounts of attorney time (and client money). But there's so much more too—think of automatic identification of terms in contracts (a problem I'm working on), or language recognition in order to build so-called "chatbots" to interact with clients. Outside of the NLP context, as I mentioned, there's facial recognition in policing; police often use other predictive models to, for example, predict likely crime locations ("heat maps") and direct more officers to the area.

There's a lot to talk about here—machine learning (a.k.a. predictive modeling, a.k.a., artificial intelligence—terminology gets a little wonky) is quickly becoming pervasive in many industries, and ours could become one of them. But I'll close with one last technical point. In a lot of ways, prediction is *easier* than inference. And the reason is that in inference you have to worry about whether there's a good causal model underlying your statistics, you have to wonder about omitted variables, you have to wonder about imagining a nonexistent effect thanks to accidentally over-interpreting p-values, all that stuff. But many of those concerns go away in the machine learning context, where you can essentially just throw massive amounts of data against the wall and try different models until one works.

You can sometimes do this in the machine learning context because, often (though not always) there's no need to *understand* the data—to know, for example, which variable in your gigantic regression is actually producing the result you see. For example, let's go back to our recurrent gender discrimination in salaries example, and the perennial problem with figuring out what to do with mediator variables like employment rank. Suppose that instead of trying to figure out whether

gender is the cause of a salary differential (the legal question) you instead just want to reliably predict people's salaries. Say you're a credit card company trying to figure out how big a credit line to give people. Well, you don't need a theory of the causal relationship between gender and rank. You can just dump everything into an extremely complicated model and then pick the parameter values that give you the best predictive power. For the purpose of guessing how much money someone will make, that works just fine.

At least, it works just fine with two caveats. First, is the danger of *overfitting*. Remember that we briefly touched on this in the context of regression. The essential problem is that the more complicated the model is, the more likely it will just generate predictions based on idiosyncratic features (noise) of the data you happen to have. But that means it won't be able to predict new, unlabeled ("out of sample") data. A model that is constrained to more resemble the actual causal processes in the world is one that will be better at predicting new data.

Second, sometimes you care about the specific causal elements in a predictive model. Our hypothetical credit card company is a good example. It would be very bad to predict salary on the basis of gender in order to set credit limits—that would, arguably, amount to gender discrimination in the credit markets. And more troublingly, it would be entirely possible that the credit card company could engage in gender discrimination even if it didn't have gender in the dataset. Indeed, we can see how that would almost certainly happen based on the problem with rank that we've discussed already! If rank is associated with salary, and gender causes rank, then predicting salary on the basis of rank will carry that influence from gender through into the predictions, and thence into the credit limits assigned by the company. (Those of you who are interested in anti-discrimination litigation: watch this space!)

That last problem is the issue of "algorithmic bias." It is particularly dangerous in a number of legal contexts. For example, suppose the police predict likelihood of being arrested in order to figure out who they should focus their attentions on. And suppose likelihood of being arrested is affected by race, because of preexisting racial biases in the criminal justice system. Then the police "heat list" just replicates existing racial injustices. That's truly extremely bad! But, according to many researchers and activists, this is a real problem that exists in real-world policing technologies. For example, there's an oft-expressed worry about California's recent replacement of cash bail with actuarial "risk assessment." In the criminal justice context, as the linked EFF statement hints, there are also substantial worries about transparency and due process, in light of the potential use of commercial models that might not be fully available to the defendant.

Algorithmic bias is an *extremely* active research area, in machine learning as well as in law. In law, one of my favorite reads on the subject is *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*, by Andrew Guthrie Ferguson. I also like Kristian Lum & William Isaac, "To Predict and Serve?" *Significance* 13(5):14-19 (2016). On the relationship between the prediction/inference dichotomy and algorithmic bias (i.e., arguments that

predictive models ought to be aimed at causal inference in order to avoid bias), see Joshua R. Loftus, Chris Russell, Matt J. Kusner, Ricardo Silva, Causal Reasoning for Algorithmic Fairness, and Chelsea Barabas, Karthik Dinakar, Joichi Ito, Madars Virza, Jonathan Zittrain, Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment. But as I said, this is a super-active research area, and the landscape changes at very high speed.

Anyway, we'll talk about more of this stuff in class. I just wanted to give you a taste.