Supervised Learning: Setting the Scene

36-290 – Statistical Research Methodology

Week 5 Tuesday – Fall 2021

The Setting

The setting for *supervised learning* is that you have a collection of $p+1$ measurements (recorded in columns of a data frame) for each of n objects (recorded in rows of a data frame). Of those measurements, p comprise the *predictor* (or *independent*) variables, with the last one being the *response* (or *dependent*) variable.

The goal: to model the data-generating process (i.e., to "learn a statistical model") and to discover associations between the predictor variables and the response variable. (While keeping in mind that "association is not causation.")

A statistical model is

$$Y|\mathbf{x} = f(\mathbf{x}) + \epsilon\,,$$

where $f(\cdot)$ is a deterministic function that represents the expected value $E[Y|\mathbf{x}]$ (the average observed value of Y, given $\mathbf{x} = \{x_1,\ldots,x_p\}$...aka "the regression line"), and ϵ is an "error" that we assume is randomly sampled from some distribution, like the normal distribution. Predicted response values are given by

$$\hat{Y}|\mathbf{x} = \hat{f}\,(\mathbf{x})$$

where the hats indicate estimated quantities.

The Overarching Question: Inference...or Prediction?

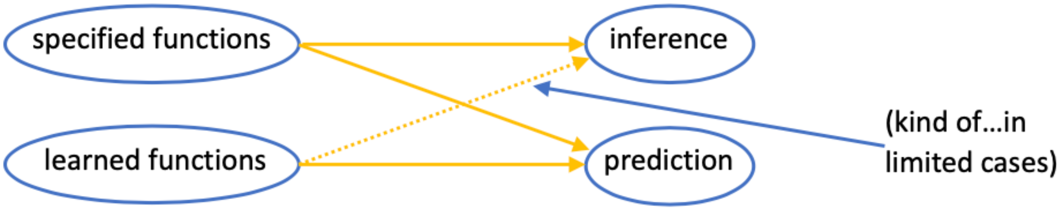
Inference: learning a statistical model and then examining and interpreting it.

- "Adding one to the number of comorbidities leads to this amount of increased cost, on average."

Prediction: learning a statistical model and then treating it as a black box.

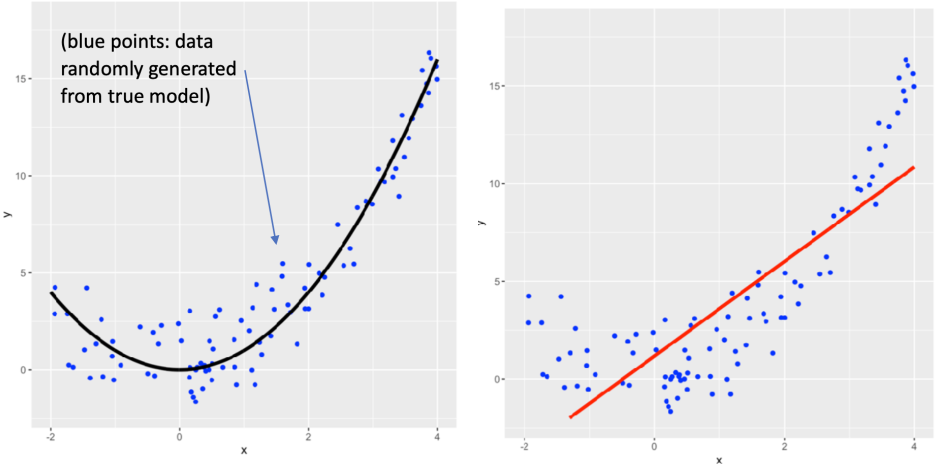
- "What is the average total insurance claim for a 60-year-old female with one intervention, no prescribed drugs, three ER visits, no complications, no comorbidities, and with a treatment duration of 100 days?"

Why does this question matter? It matters because it impacts the possible suite of models that you can apply to your dataset.



- "specified functions": functions you can write down yourself, on paper (e.g., linear regression)
- "learned functions": functions a machine learns via algorithm, given data (e.g., random forest)

Inference vs. Prediction: the Tradeoff


(blue points: data randomly generated from true model)

As shown in the left panel, data (blue points) are generated from a smoothly varying non-linear model (black line), with random noise added. In the right panel, the red line shows a simple linear regression fit to the observed data.

The statistical model of linear regression is fully parameterized and completely inflexible. As you can see, it does not provide a good estimate of $f(\mathbf{x})$…but it has the virtue of being easy to interpret.

The basic tradeoff: the more flexible a model is, the better predictions it will generate, but the harder it will be to explain.

It is very important to determine, at the start of any analysis, whether the goal is inference or prediction, and if it is inference, how important inference is, and how much inferential ability you are willing to give up if predictive models provide substantially better fits to your data.

Inference vs. Prediction: Two More Important Points

- If inference is your goal, you should still always include "learned function" models like random forest in your suite of models that you will apply in an analysis.

Why? If you do not try such nonlinear models, you will never know how much better such models might perform…this information is critical. If inference is your goal, and the MSE for a random forest model is half that seen with linear regression, then you know that any inferences you draw from the latter model may not reflect reality (i.e., your inferences may be afflicted by biases: remember, inflexible models are the high-bias models, assuming that the true association is not linear).

- If prediction is your goal, you should still always include "specified function" models like linear regression in your suite of models that you will apply in an analysis.

Why? For the simple reason that the true association between the predictors and the response could be (approximately) linear. If this is the case, machine learning will not help you…and you will get the best of both worlds (inferential ability *and* predictive ability).

In short: try all reasonable models, regardless of the analysis goal. Then sort out the results.