D.1: Model Interpretability

Dr. Bean - Stat 5100

In what ways is a linear model "interpretable"? In what ways is it NOT interpretable?

Interpretable:

- Relatively easy to explain how the model works.
- Each explanatory variable is associated with a single number that we can directly interpret (assuming no interactions).

NOT Interpretable:

- Linear models are often heavily manipulated "by hand", making it hard to separate what is a product of the data, from what is a product of our manipulations.
- If our associations are plagued by **confounding variables** or the data itself contains bias or prejudice, we may make inappropriate conclusions from our model, regardless of the satisfaction of assumptions.

Why should we care about interpretability?

Consider the European Unions Law requiring a "right to explanation" (`https://arxiv.org/pdf/1606.08813.pdf`). As more and more significant decisions begin being made by algorithms, more and more people will want to know why and how those decisions are being made.

Example: a person would probably like to know why a statistical algorithm denied them the opportunity for a home loan.

How might a statistical model be racist/prejudiced? Are linear models also prone to prejudice? If so, how?

If we use historical data to train our models (say loan default risk), and those historical decisions were made in a prejudiced way, the model will likewise learn to be prejudiced. Also, confounding variables (an unaccounted for variable that is highly correlated with two other variables and makes the two other variables look correlated with each other) make can make things like "race" or "gender" drive model predictions, even if those variables are actually included in the model.

How might statistics be used to ensure that the predictive models of tomorrow don't suffer from the prejudices of yesterday?

There is no right answer here. However, a better understanding of how our predictive models work will help us know how and when the results might betray us.