Modeling and Bias

Clark

Modeling

Last lesson, you discussed the movie Coded Bias. I recently read a review of this that states ''Eye-opening Netflix doc faces racist technology" Today, I want to talk a little about the math behind the algorithms and for us to think through is the technology racist? Or is there something else entirely going on here.

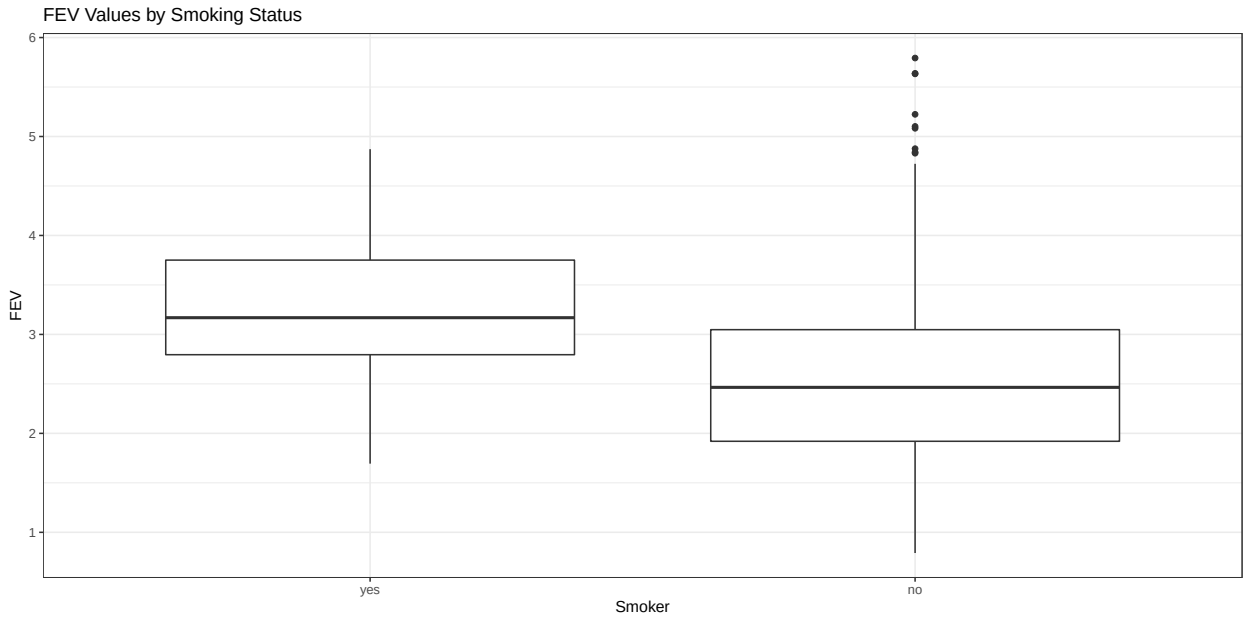
To start, I want us to consider the point of building statistical models. In your opinion, when we collect data and build out a model, what are the general tasks we are trying to accomplish?

I want to pull on a thread here comparing two of what, I view, are potential goals. Descriptive vs diagnostic. In general, I think of a (parametric) model as doing the following:

So, as we collect data, we kind of move up this tree. We have the goals of either understanding the state of the world, or predicting why the world is in the state that it is. Let's take an example that you may have seen before.

```
FEV <- read.delim("FEV.txt")
FEV <- FEV %>%
  mutate(Gender=factor(Gender, levels=c('Male', 'Female')),
         Smoker = factor(Smoker, levels=c('yes','no')))
```

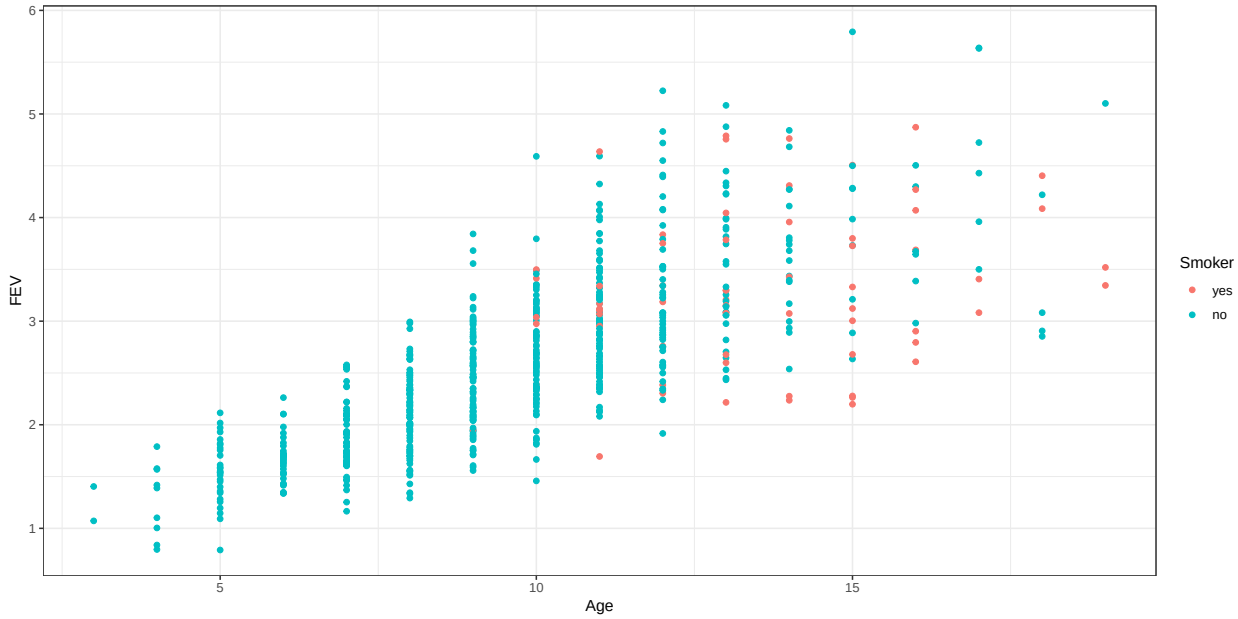
```
ggplot(FEV,aes(x=Smoker, y=FEV))+
  geom_boxplot()+
  theme_bw()+
  ggtitle("FEV Values by Smoking Status")
```



So, underlying this is a model. The model says that there is a relationship between FEV and smoking status. But is it surprising?

The issue here is, in our minds, we are making a diagnostic statement, but the model is only a descriptive model. It cannot tell us 'why' a phenomenon is occurring it can only tell us what is occurring.

```
ggplot(FEV,aes(x=Age, y=FEV, color=Smoker)) +
  geom_point() +
  theme_bw()
```



The issue here, isn't the data, nor is it the algorithm, it is:

Models can do what they were built to do. A neural network is a model.

Let's think through what purpose this model was likely built for.

Bias

Now, in the documentary the models weren't necessarily being used inappropriately, but there still was an issue. We have this word *bias*. But what does it mean? Take a few minutes and brain storm what you think the word bias means.

For me it has a very clear meaning.

$$E[\hat{\theta}] - \theta$$

Let's talk through this a bit.

So, how does this work? Well, let's consider the following example. Let's assume that we want to know the average height of Cadets in the corps. Here θ represents the true, unknown, average height. So, I go and I sample a bunch of cadets randomly and I get:

```
heights <- c(67,69,72, 65,69,70,71,67)
```

Come up with two different ways we could estimate θ

Why might we prefer using $\bar{x}$?

What if the data looked like:

```
heights <- c(84,69,72, 65,69,70,71,67)
```

So, sometimes, it might make sense to use a biased sampling method. Here, we say that the *method* we are using to make inference about θ is *biased*.

In fact, may of the classic data science algorithms are biased on purpose! Why might we want to do that? Well, it turns out that a good model, perhaps, shouldn't just be scored on *bias* but perhaps should be scored on mean squared error:

$$E[(\hat{\theta} - \theta)^2]$$

Which, we can decompose like:

So, oftentimes data scientists will use methods that have bias, but have a low variance. Meaning, we would prefer the following marksman:

So, certainly our methodologies can create bias, but our *data* can also be biased. Let's consider the following. We want to determine the height of the corps, so we walk into a class that is entirely full of males and calculate:

```
male_heights <- c(67,69,72, 72,69,70,71,67)
mean(male_heights)
```

```
## [1] 69.625
```

Does $E[\hat{\theta}] - \theta = 0$ here? What is the issue?

Herein likes the big issue in data science today. *Our sample is not representative of the population we are trying to say/predict something about.*

What does it mean to have a representative sample?

In traditional statistics this would be handled through experimental design. However, most people don't study this anymore...

Note that sometimes we think this can be handled by increasing our sample size. Why is this an issue?

So, how do we get over this?

Let's take a look at https://news.mit.edu/2022/machine-learning-biased-data-0221

What are the authors arguing here?

Another possibility is through doing a different sort of model fitting that allows us to inject our own knowledge into the estimate. Say, for instance, we know that we have a biased dataset and we cannot get beyond this. If we just had males heights and we wanted to estimate heights for our entire population what could we do?

Related to this is *Bayesian inference*.

To think through how this work, let's consider the case where we are trying to determine the probability of a coin being fair, so we want to estimate the probability of obtaining a Heads. If we flip a coin five times and get all heads. Using what we were taught in MA206, what would the probability of obtaining a heads be? Is this logical?

Neural Networks

So, let's bring this all back around. A neural network, at the root of it, is nothing more than a statistical model. While we may see it written as a set of connected nodes and edges, we can actually express (at least for a single layer model) as a non-linear regression model.

So, the weights are nothing more than parameters that we have to estimate. How do we estimate them then? Well, same way we estimate everything in statistics. Through minimizing a loss function...

Now, there's some nuance here. Since we generally have a model that has more parameters than data points, we can't actually fit it, so we stop early to prevent 'overfitting'.

The point is this. The model actually can be fit really easily (code wise). https://datascienceplus.com/fitting-neural-network-in-r/

I don't need to know a SINGLE thing about the model or how it is being fit to go about and fit it. Our computer systems are complex enough that we can fit pretty sophisticated models with little effort. So, whose responsibility is it if we misuse the tools? Is it the technology? Is it the person who federated out the algorithms? Is it the person who collected the data?