Categorical Variables

DASC 512

Categorical (Qualitative) Variables

In order to include categorical independent variables, we must code them numerically by using dummy (indicator) variables valued 0 or 1

We'll use one fewer dummy variable than the number of levels

- For example, if we have three categories: "Red" "White" "Blue", then

	x_1	x_2
Red	0	0
White	1	0
Blue	0	1

- x_1: Color is white, x_2: Color is Blue

Example: IQ tests

Lucky for us, Python does this automatically for regression

But, **remember this concept** because you'll have to apply it for the machine learning algorithms you'll use next quarter.

Let's look at an example using IQ tests and gender to predict test scores

Example: FEV response by drug treatment

Remember this one from the Progress Checks?

We did ANOVA on it before. Let’s look at it as a regression model.

Recap and Review

- Whenever the relationship between an x and y does not appear linear, we can add higher order terms
- When one variable has an effect on the relationship between another variable and y, we can add interaction terms
- Whenever interactions or higher-order teams are included, the first-order terms must also be included.
- Using the C() wrapper in Patsy will create dummy variables for you

Next time…

Model Building