Lesson 6

Linear Regression

y = beta * X + alpha (+ error)

The relationship between the matrix **X** and a dependent vector **y** using a y-intercept **alpha** and the relative coefficients **beta**

Linear regression works best when:

- The data is normally distributed (but doesn't have to be)
- X's significantly explain y (have low p-values)
- X's are independent of each other (low multicollinearity)
- Resulting values pass linear assumption (depends upon problem)
- If data is not normally distributed, we could introduce *bias***.**

Sklearn:

- Models are defined as objects
- All sklearn modeling classes are based on the base estimator. This means all models take a similar form.
- All estimators take a matrix **X**, either sparse or dense.
- Supervised estimators also take a vector **y** (the response).
- Estimators can be customized through setting the appropriate parameters.

Classes are an abstraction for a complex set of ideas, e.g. *human.*

Specific **instances** of classes can be created as **objects**.
‣*john_smith = human()*

Objects have **properties**. These are attributes or other information.
‣*john_smith.age*
‣*john_smith.gender*

Object have **methods**. These are procedures associated with a class/object.
‣*john_smith.breathe()* ‣*john_smith.walk()*

Multiple regression analysis uses multiple variables to predict a dependent variable. They need to be mostly independent to avoid multicollinearity - if two or more variables are highly correlated then this can cause problems with the model.

Dummy variables can be created for categorical variables in order to create a regression model.