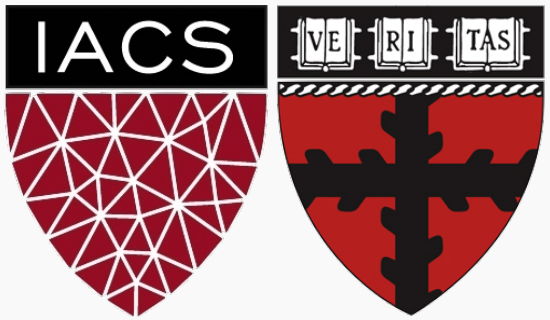
Lecture #4: Introduction to Regression

Data Science 1
CS 109A, STAT 121A, AC 209A, E-109A

Pavlos Protopapas Kevin Rader
Margo Levine Rahul Dave



Lecture Outline

Announcements

Data

Statistical Modeling

Regression vs. Classification

Error, Loss Functions

Model I: k-Nearest Neighbors

Model II: Linear Regression

Evaluating Model

Comparison of Two Models

Statistical Modeling

Predicting a Variable

Let's image a scenario where we'd like to predict one variable using another (or a set of other) variables.

Examples:

- Predicting the amount of view a YouTube video will get next week based on video length, the date it was posted, previous number of views, etc.
- Predicting which movies a Netflix user will rate highly based on their previous movie ratings, demographic data etc.
- Predicting the expected cab fare in New York City based on time of year, location of pickup, weather conditions etc.

Outcome vs. Predictor Variables

There is an asymmetry in many of these problems: the variable we'd like to predict may be more difficult to measure, is more important than the other(s), or may be directly or indirectly influenced by the values of the other variable(s).

Thus, we'd like to define two categories of variables: variables whose value we want to predict and variables whose values we use to make our prediction.

Outcome vs. Predictor Variables

Definition

Suppose we are observing $p + 1$ number variables and we are making n sets observations. We call

- the variable we'd like to predict the ***outcome*** or ***response variable***; typically, we denote this variable by Y and the individual measurements y_i.
- the variables we use in making the predictions the ***features*** or ***predictor variables***; typically, we denote these variables by $X = (X_1, \ldots, X_p)$ and the individual measurements $x_{i,j}$.

Note: i indexes the observation ($i = 1, 2, \ldots, n$) and j indexes the value of the j-th predictor variable ($j = 1, 2, \ldots, p$).

True vs. Statistical Model

We will assume that the response variable, Y, relates to the predictors, X, through some unknown function expressed generally as:

$$Y = f(X) + \epsilon.$$

Here,

- f is the unknown function expressing an underlying rule for relating Y to X,
- ϵ is random amount (unrelated to X) that Y differs from the rule $f(X)$

A ***statistical model*** is any algorithm that estimates f. We denote the estimated function as $\widehat{f}$.

Prediction vs. Estimation

For some problems, what's important is obtaining $\widehat{f}$, our estimate of f. These are called ***inference*** problems.

When we use a set of measurements of predictors, $(x_{i,1}, \ldots, x_{i,p})$, in an observation to predict a value for the response variable, we denote the ***predicted value*** by $\widehat{y}_i$,

$$\widehat{y}_i = \widehat{f}(x_{i,1}, \ldots, x_{i,p}).$$

For some problems, we don't care about the specific form $\widehat{f}$, we just want to make our prediction $\widehat{y}_i$ as close to the observed value y_i as possible. These are called ***prediction problems***.

We'll see that some algorithms are better suited for inference and others for prediction.

Regression vs. Classification

Outcome Variables

There are two main types of prediction problems we will see this semester:

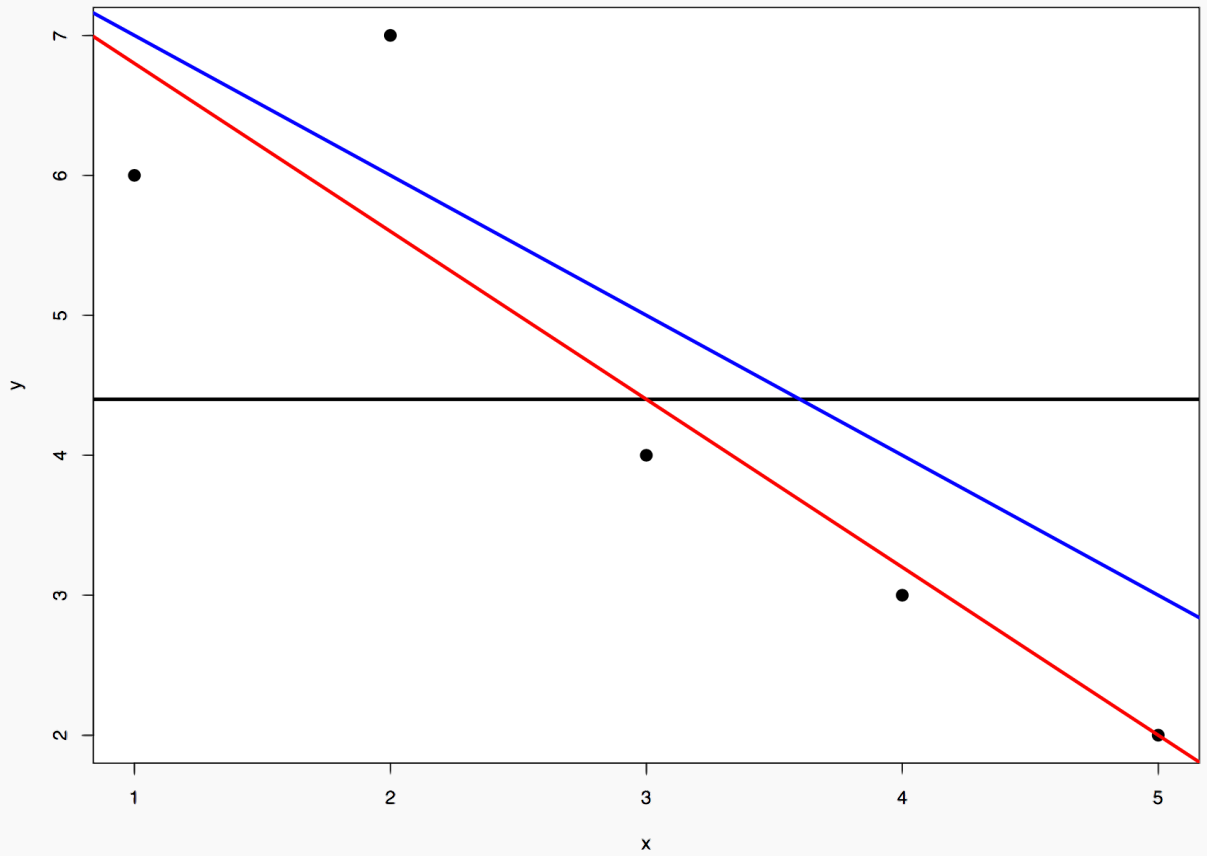
- ***Regression problems*** are ones with a quantitative response variable.
  Example: Predicting the number of taxicab pick-ups in New York.
- ***Classification problems*** are ones with a categorical response variable.
  Example: Predicting whether or not a Netflix user will like a particular movie.

This distinction is important, as each type of problem may require it's own specialized algorithms along with metrics measuring effectiveness.

Error, Loss Functions

Line of Best Fit

Which of the following linear models is the best? How do you know?



Using Loss Functions

Loss functions are used to choose a suitable estimate $\widehat{f}$ of f.

A statistical modeling approach is often an algorithm that:

- assumes some mathematical form for f, and hence for $\widehat{f}$,
- then chooses values for the unknown parameters of $\widehat{f}$ so that the loss function is minimized on the set of observations

Error & Loss Functions

In order to quantify how well a model performs, we define a ***loss*** or ***error function***.

A common loss function for quantitative outcomes is the ***Mean Squared Error (MSE)***:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

The quantity $|y_i - \widehat{y}_i|$ is called a ***residual*** and measures the error at the i-th prediction.

Caution: The MSE is by no means the only valid (or the best) loss function!

Question: What would be an intuitive loss function for predicting categorical outcomes?