DATA 8

Fall 2016

Lecture 30, November 4

Least Squares

Slides created by Ani Adhikari and John DeNero

Announcements

- Project 2 deadline is Tuesday 11/8 at 7 p.m.
- Homework will be assigned today

- Only 10 lectures till RRR week!

Regression to the Mean

- **estimate of $y = r \cdot x$**, when both variables are measured in standard units
- On average (though not for each individual), regression predicts *y* to be closer to the mean than *x* is

Slope, Intercept, and Fitted Values

"fitted value" of y = slope * x + intercept

where

$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

(Demo)

Units of the Slope

units of *y* per unit of *x*

- Predicting weight based on height
  - slope 3.6 pounds per inch
- If Person A is 1 inch taller than Person B, then predict Person A to be 3.6 pounds heavier than Person B
- If two groups are 1 inch apart in height, then the average weight of the taller group is about 3.6 pounds more than the average weight of the shorter group.

Error in Estimation

- **error = actual value - estimate**
- Typically, some errors are positive and some negative
- To measure the rough size of the errors
  - **square** the **errors** to eliminate cancellation
  - take the **mean** of the squared errors
  - take the square **root** to fix the units
  - **root mean square erro**r (rmse)

(Demo)

Least Squares Line

- Minimizes the root mean squared error (rmse) among all lines
- Equivalently, minimizes the mean squared error (mse) among all lines
- Names:
  - “Best fit” line
  - Least squares line
  - Regression line

(Demo)

Numerical Optimization

- Numerical minimization is approximate but effective
- Lots of machine learning involves numerical minimizing error
- If the function **`mse(a, b)`** returns the mse of estimation using the line "estimate = a*x* + b",
  - then **`minimize(mse)`** returns array $[a_0, b_0]$
  - a_0 is the slope and b_0 the intercept of the line that minimizes the mse among lines with arbitrary slope `a` and arbitrary intercept `b` (that is, among all lines)