DATA 8
Summer 2018

Lecture 27

Linear Regression

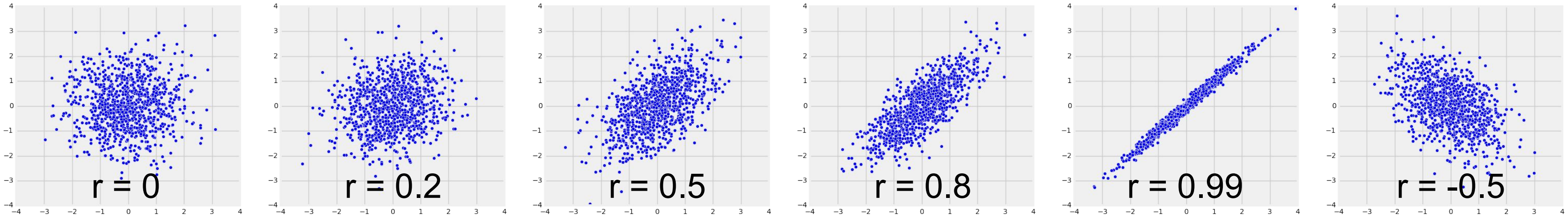
Slides created by John DeNero (denero@berkeley.edu) and Ani Adhikari (adhikari@berkeley.edu)
Contributions by Fahad Kamran (fhdkmrn@berkeley.edu) and Vinitra Swamy (vinitra@berkeley.edu)

Announcements

Correlation (Review)

The Correlation Coefficient *r*

- Measures ***linear*** association
- Based on standard units
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; *uncorrelated*



Definition of *r*

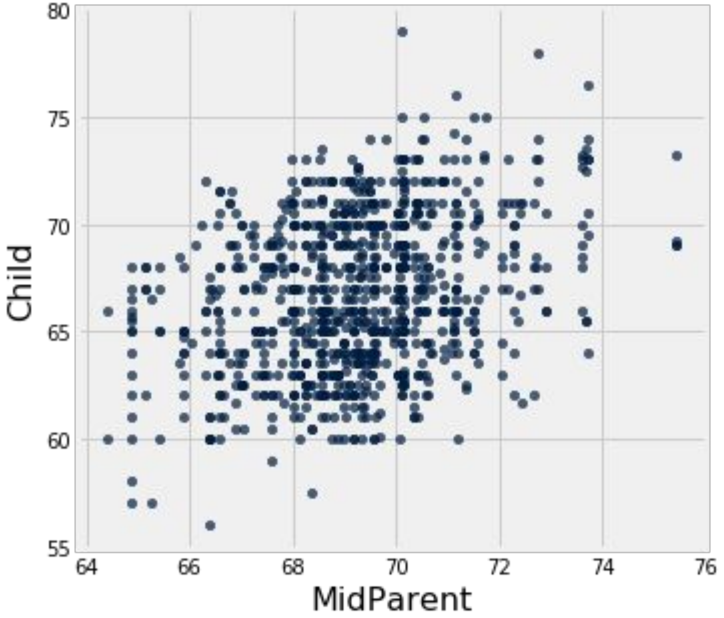
Correlation Coefficient (*r*) =

average of	product of	x in standard units	and	y in standard units

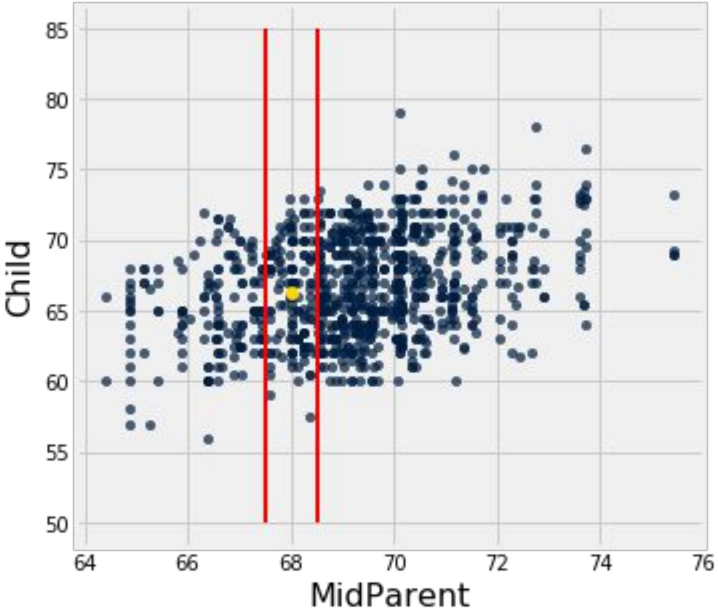
Measures how clustered the scatter is around a straight line

Prediction

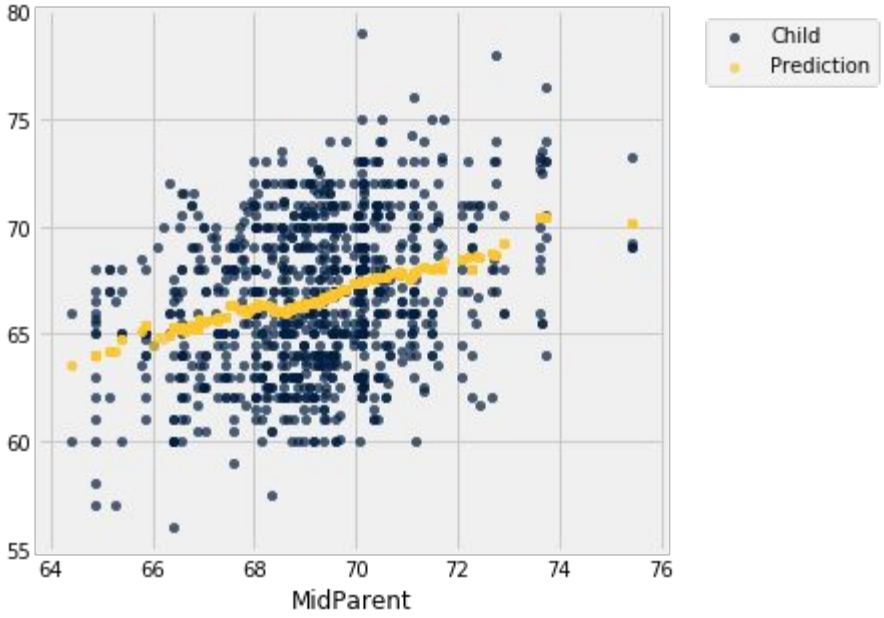
Galton's Heights



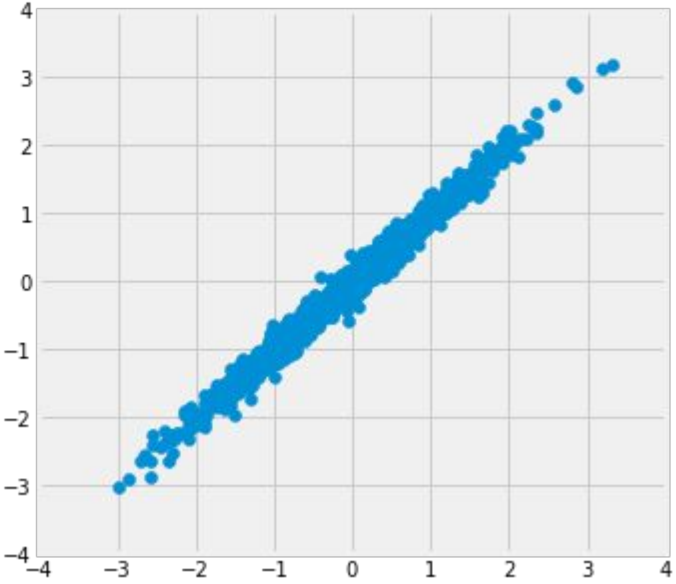
Galton's Heights



Galton's Heights

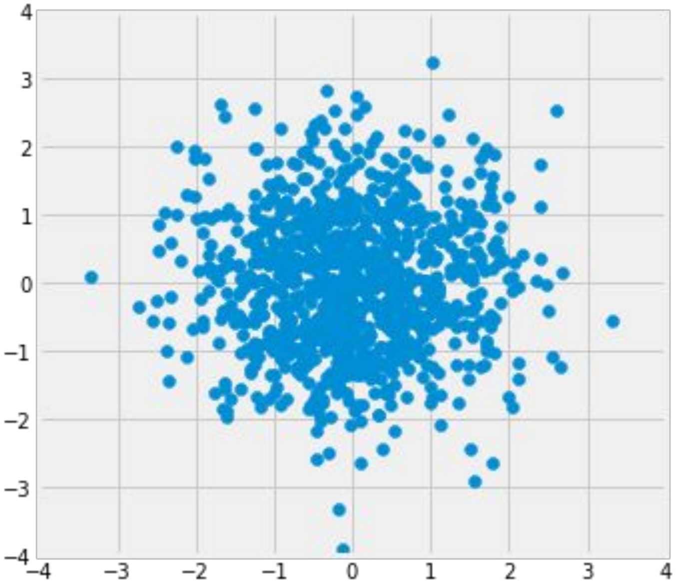


Where is the prediction line?



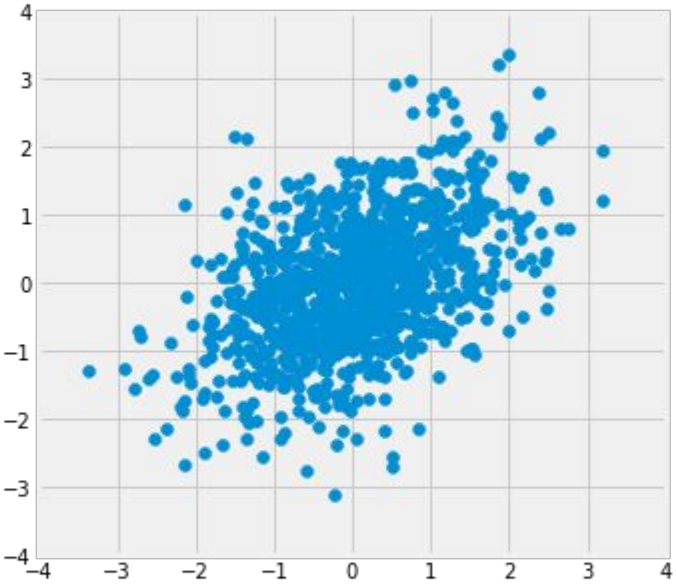
r = 0.99

Where is the prediction line?



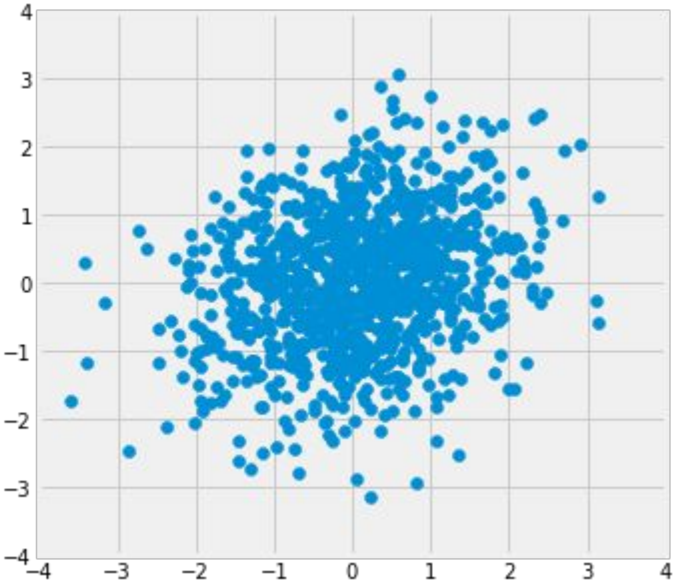
r = 0.0

Where is the prediction line?



r = 0.5

Where is the prediction line?



r = 0.2

Nearest Neighbor Regression

A method for prediction:

- Group each x with a representative x value (rounding)
- Average the corresponding y values for each group

For each representative x value, the corresponding prediction is the average of the y values in the group.

Graph these predictions.

If the association between x and y is linear, then points in the graph of averages tend to fall on the regression line.

Linear Regression

(Demo)

Regression to the Mean

A statement about x and y pairs

- Measured in *standard units*
- Describing the deviation of x from 0 (the average of x's)
- And the deviation of y from 0 (the average of y's)

On average, y deviates from 0 less than x deviates from 0

Regression Line

$$\boxed{y_{(\mathrm{su})} = \underbrace{r}_{\text{Correlation}} \times x_{(\mathrm{su})}}$$

Not true for all points — a statement about averages

Slope & Intercept

Regression Line Equation

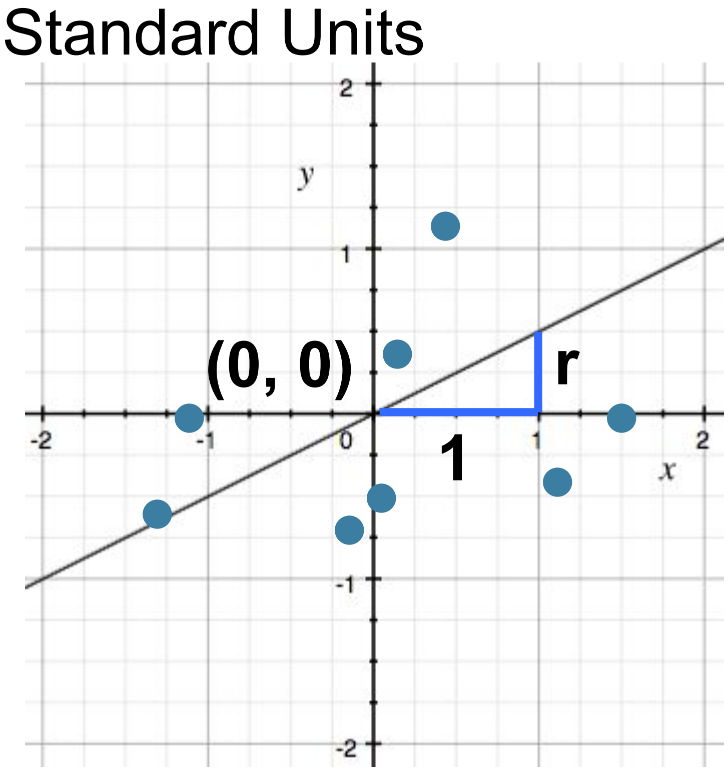
In original units, the regression line has this equation:

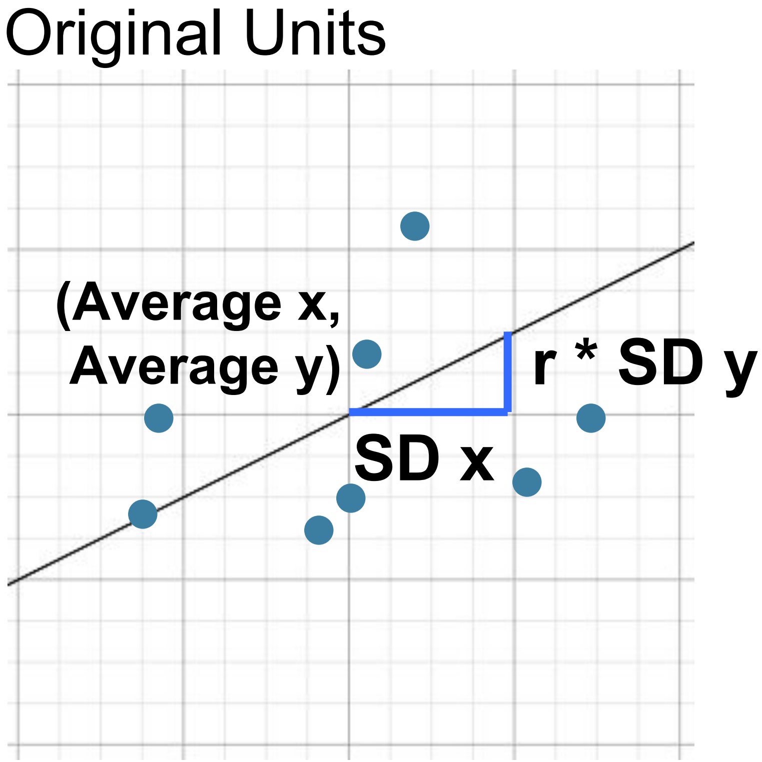
$$\underbrace{\frac{\text{estimate of } y \; - \; \text{average of } y}{\text{SD of } y}}_{\text{estimated y in standard units}} = r \times \underbrace{\frac{\text{the given } x \; - \; \text{average of } x}{\text{SD of } x}}_{\text{x in standard units}}$$

Lines can be expressed by *slope* & *intercept*

$$y = \text{slope} \times x + \text{intercept}$$

Regression Line





Slope and Intercept

estimate of y = slope * x + intercept

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

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

(Demo)

Scenario Question

You use a regression line to predict height based on weight and get a slope of .52 inches per pound.

I eat a lot of ice cream and gain 1 pound.

True or False:

> My regression line predicts I will gain 1 inch on my height

Scenario Question

False

The regression line is a **statement about averages**. Given two groups of people with 1 lb difference, we expect the average height of the heavier group to be .52 inches greater than the average height of the lighter group

The regression line is based on a snapshot of time and looking at holistic trends. We are **not** following 1 person and intending to make a prediction about them