YData: An Introduction to Data Science

Lecture 30: Linear Regression

Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

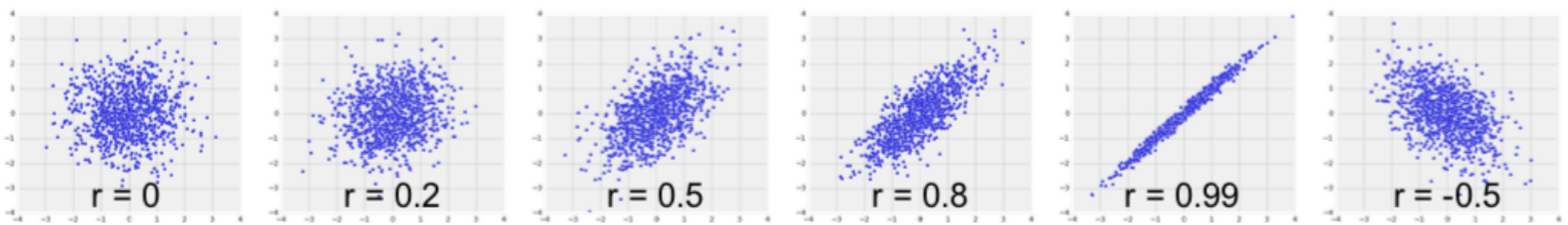
Reminders

- Assignment 09 due Thursday 4/15
- Project 2 due on Friday 4/16
- Second-to-last assignment released Friday

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*

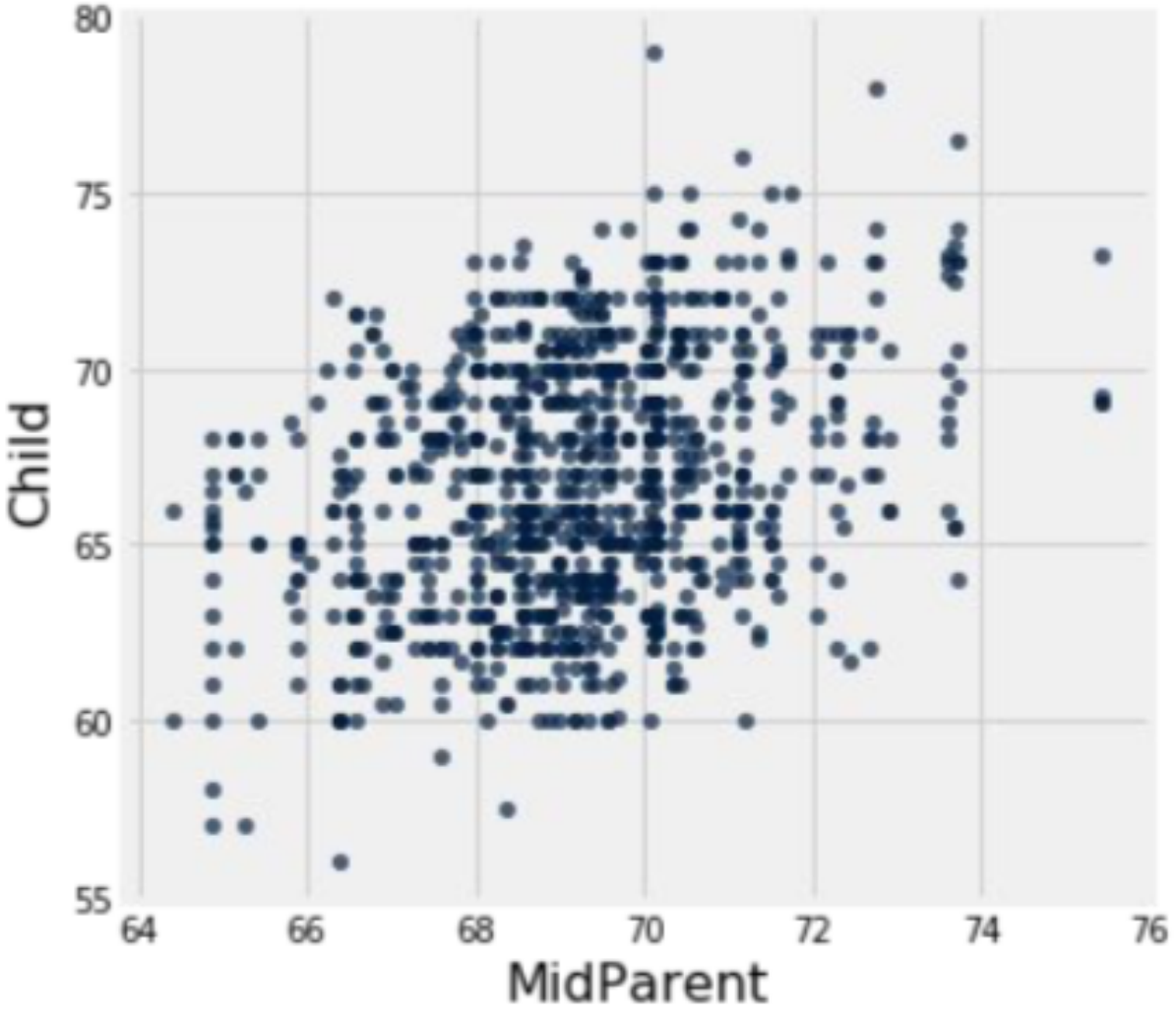
Correlation Coefficient (r) =

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

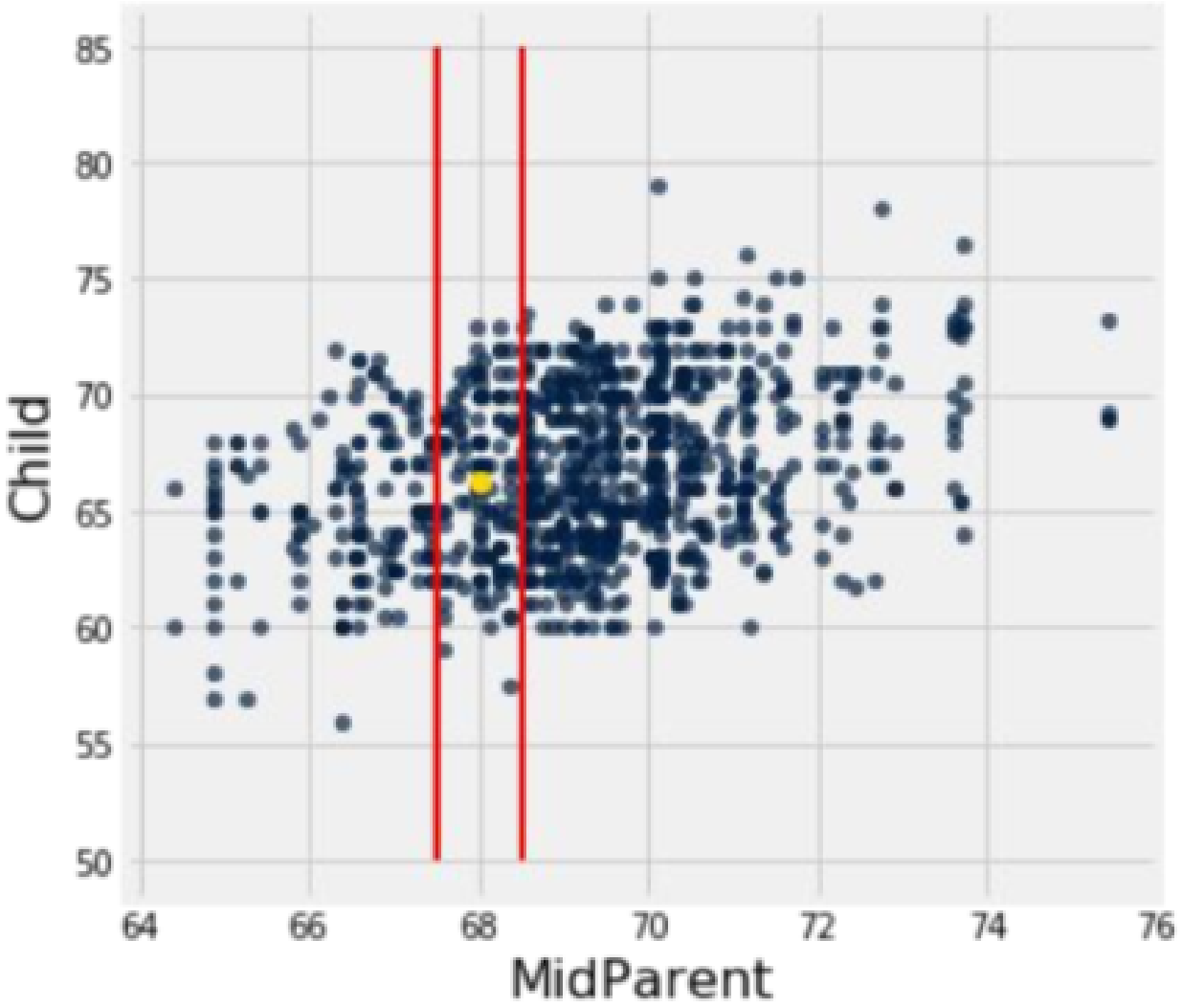
Measures how clustered the scatter is around a straight line

Prediction

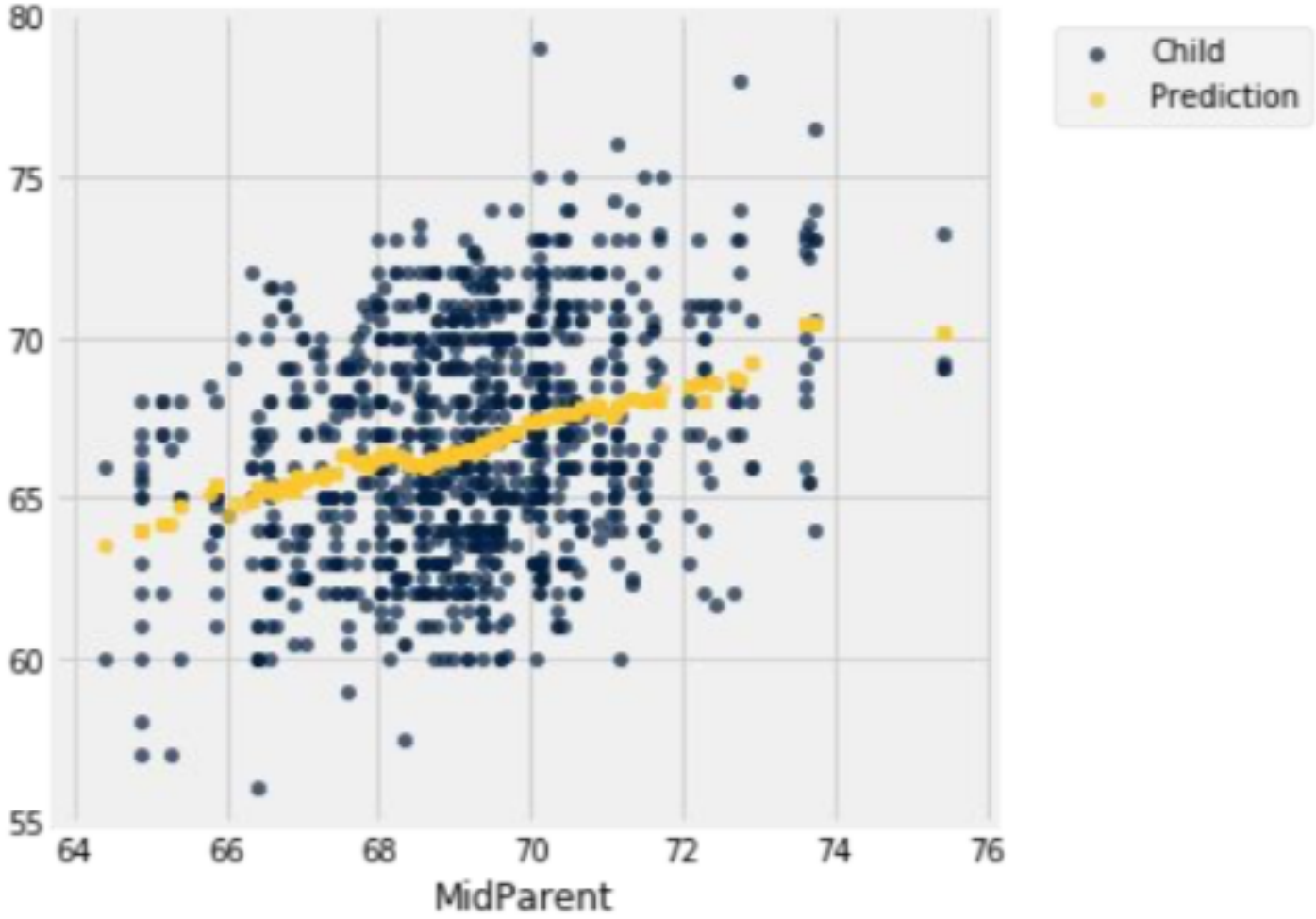
Galton's Heights



Galton's Heights



Galton's Heights



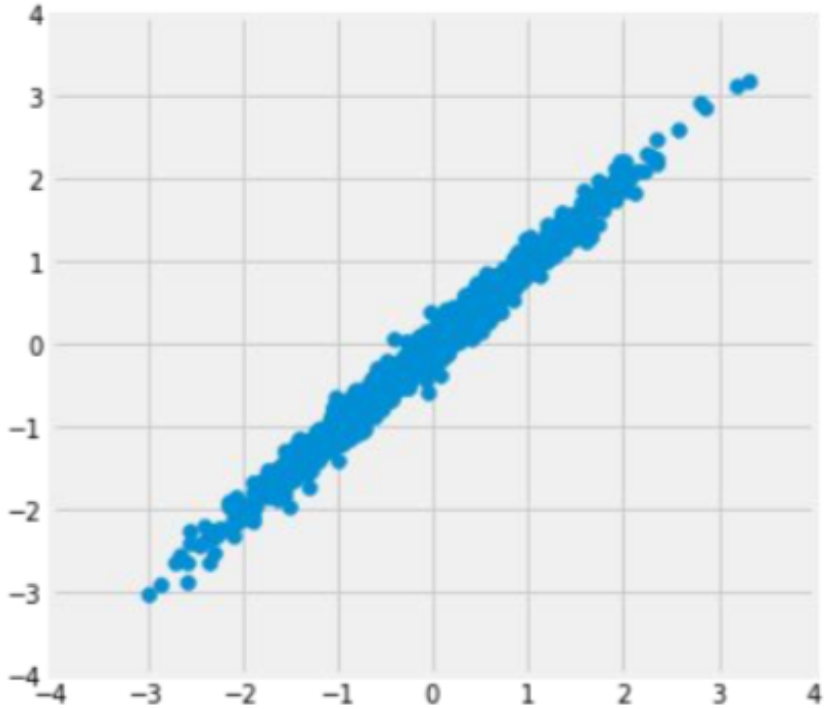
Prediction and Correlation

Today we're going to connect prediction and correlation.

Let's first review some properties of correlation through more examples.

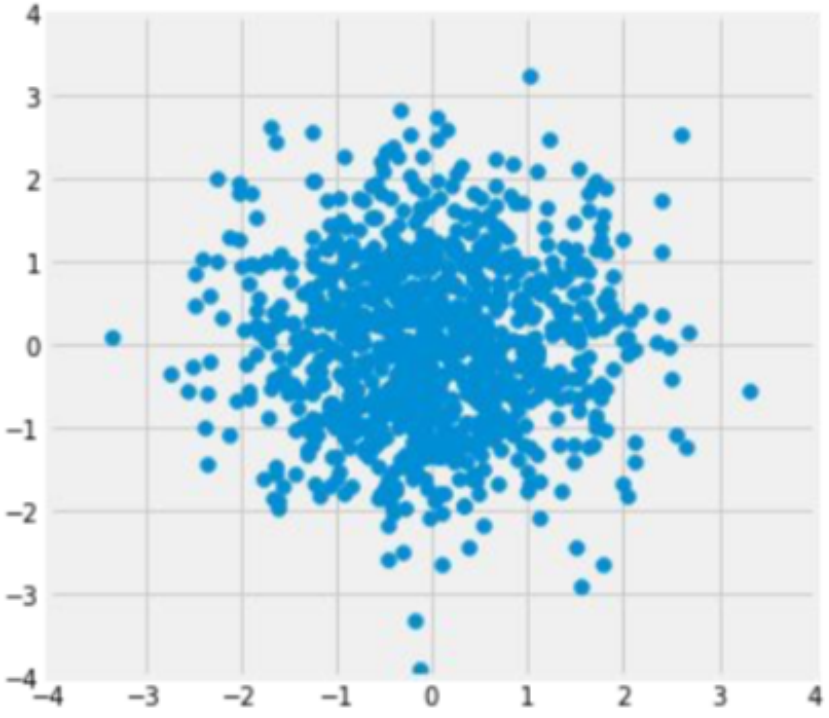
(DEMO)

Where is the prediction line?



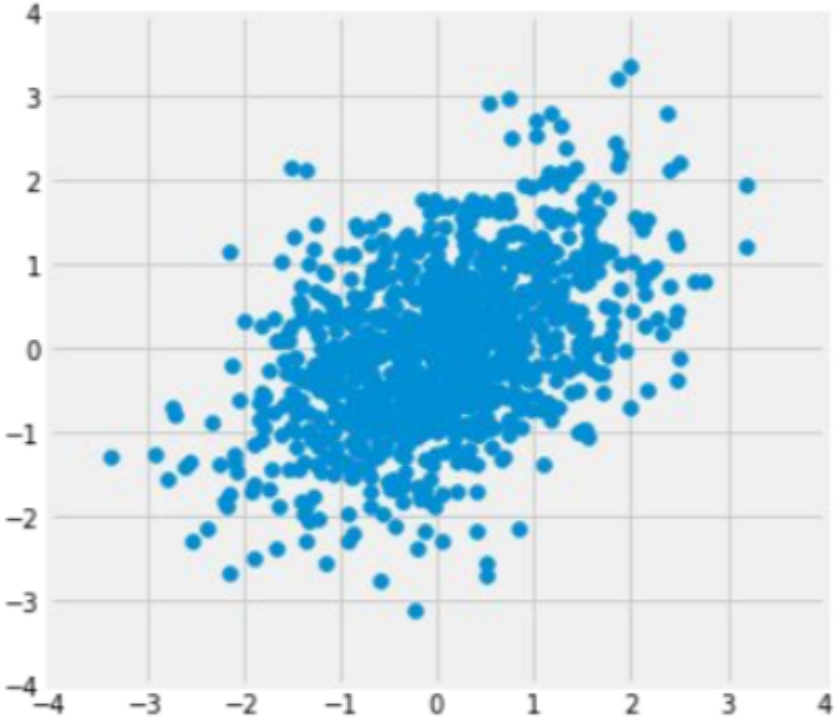
$r = 0.99$

Where is the prediction line?



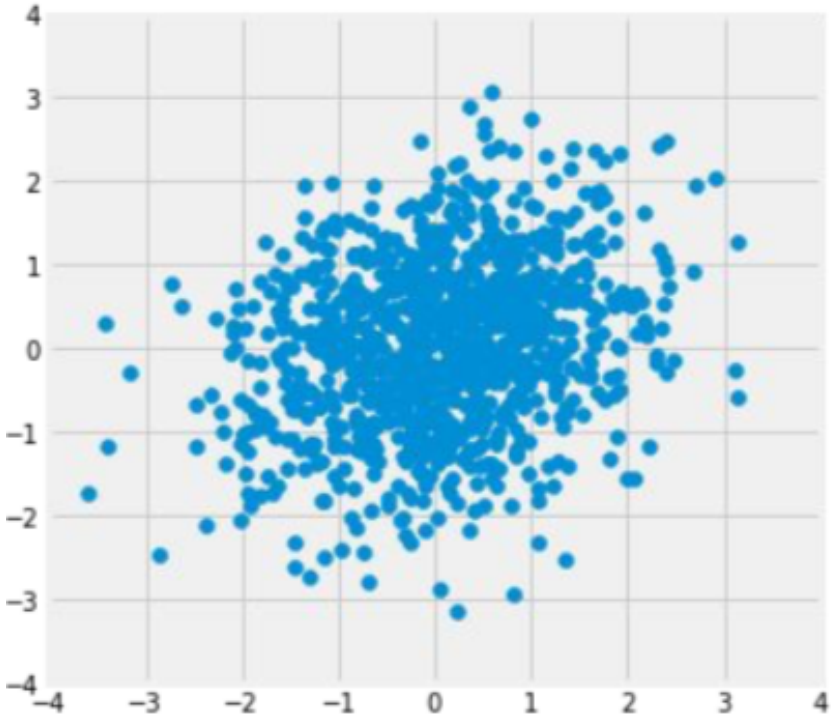
$r = 0.0$

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

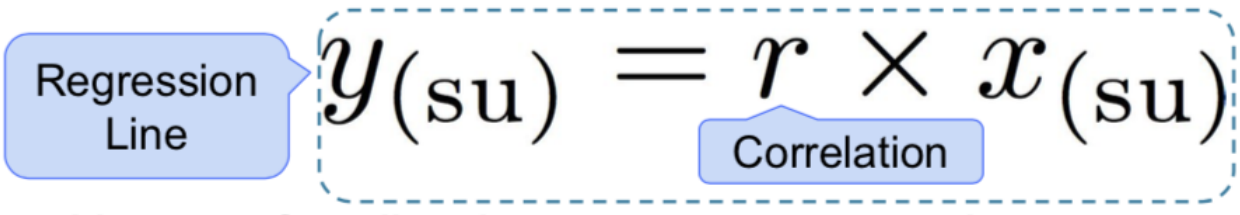
(DEMO)

Regression to the Mean

A statement about x and y pairs

- Measured in *standard units*
- Describing the deviation of y's from their mean (the average of the y's) for a fixed x.

On average, y deviates around it's mean for a given x less than x deviates from 0



So, the *average* y value for a given x (in standard units) is $r \times x$.

The Regression Effect

- It's a statement about averages
- Example: Take all children whose midparent height is 1.5 standard unit. The average height of these children is somewhat *less* than 1.5 standard units.
- It doesn't say that all of these children will be somewhat less than 1.5 standard units in height. Some will be taller, and some will be shorter.

Slope & Intercept

Regression Line Equation

In original units, the regression line has this equation:

$$\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}$$

estimated y in standard units

x in standard units

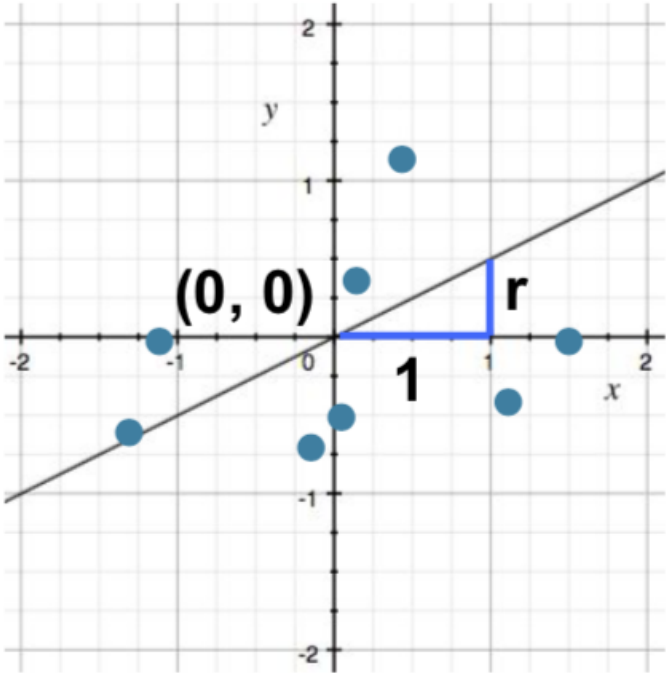
Lines can be expressed by slope & intercept

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

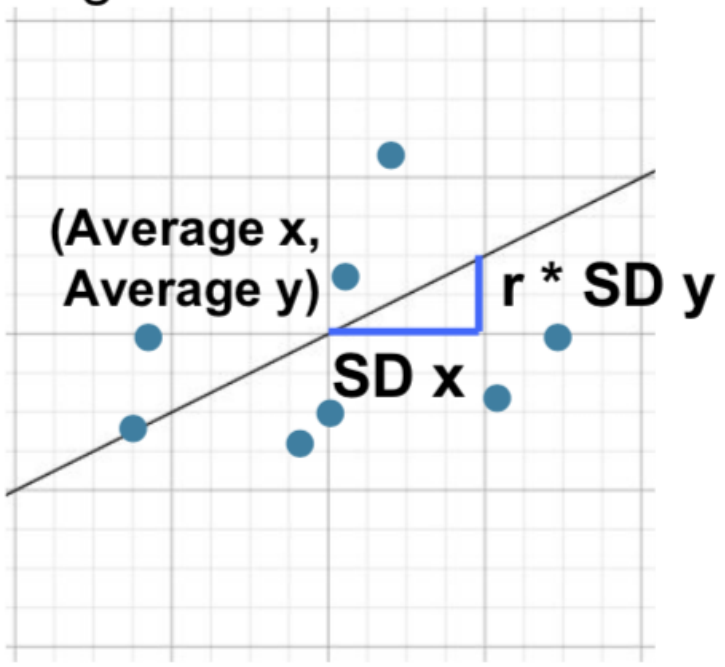
With a little algebra, we can calculate the slope and intercept

Regression Line

Standard Units



Original Units



Slope and Intercept

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

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

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

(DEMO)