YData: An Introduction to Data Science

Lecture 30: Linear Regression

Jessi Cisewski-Kehe and John Lafferty
Statistics & Data Science, Yale University
Spring 2019

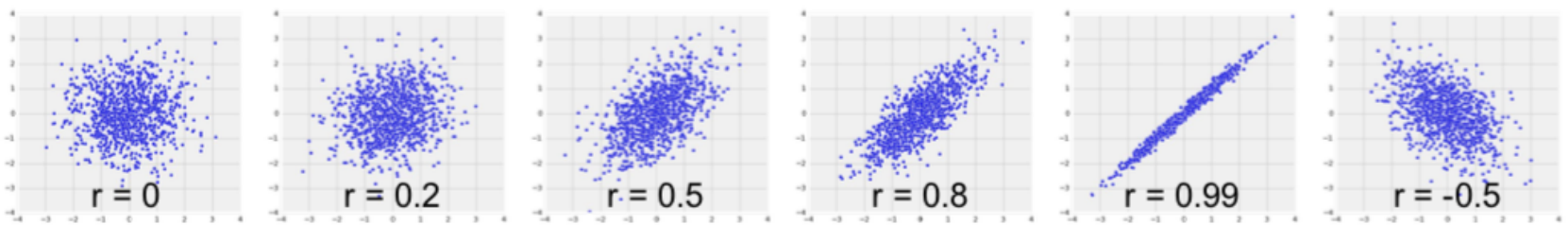
Credit: data8.org

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*

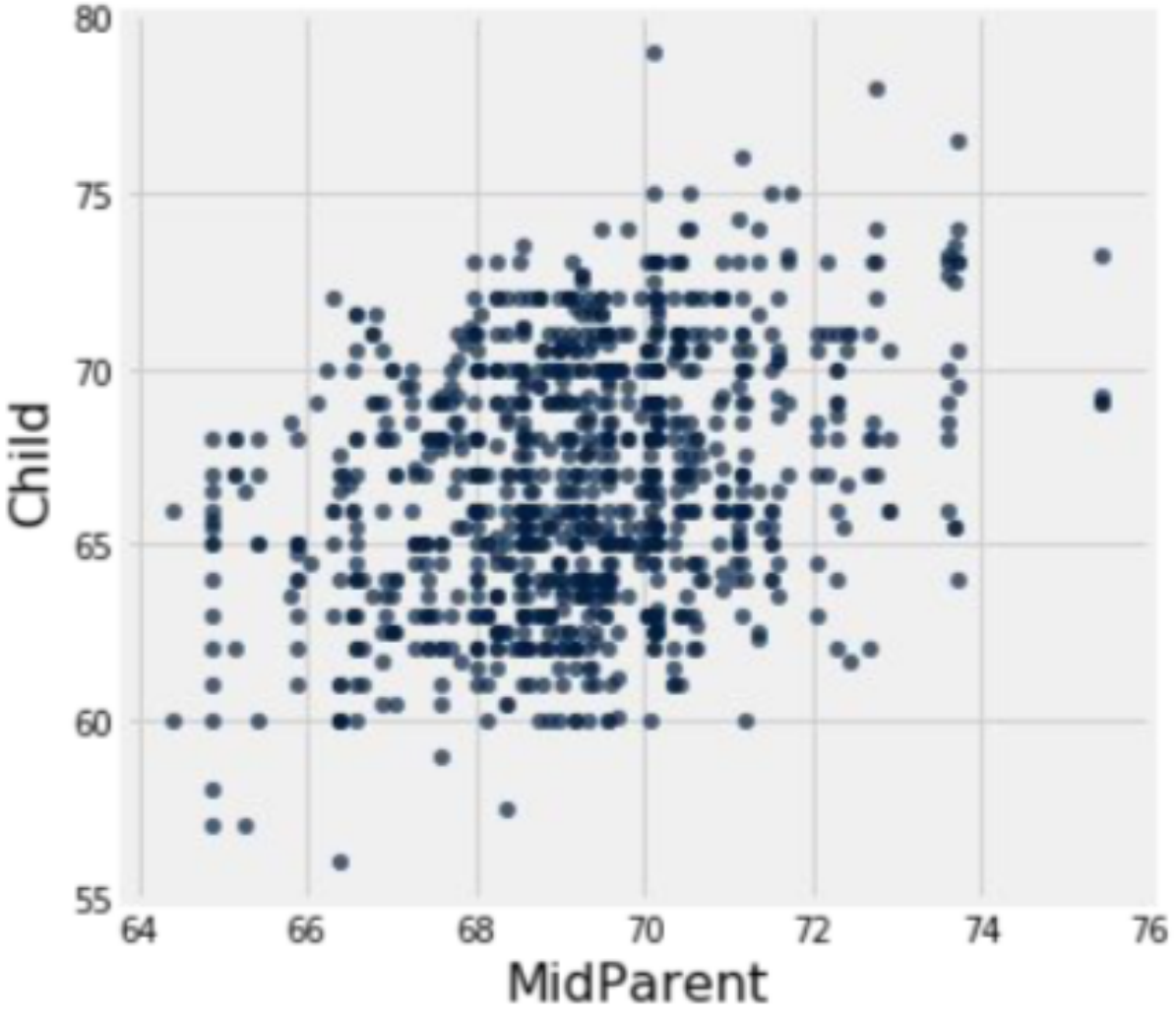
Correlation Coefficient (r) =

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

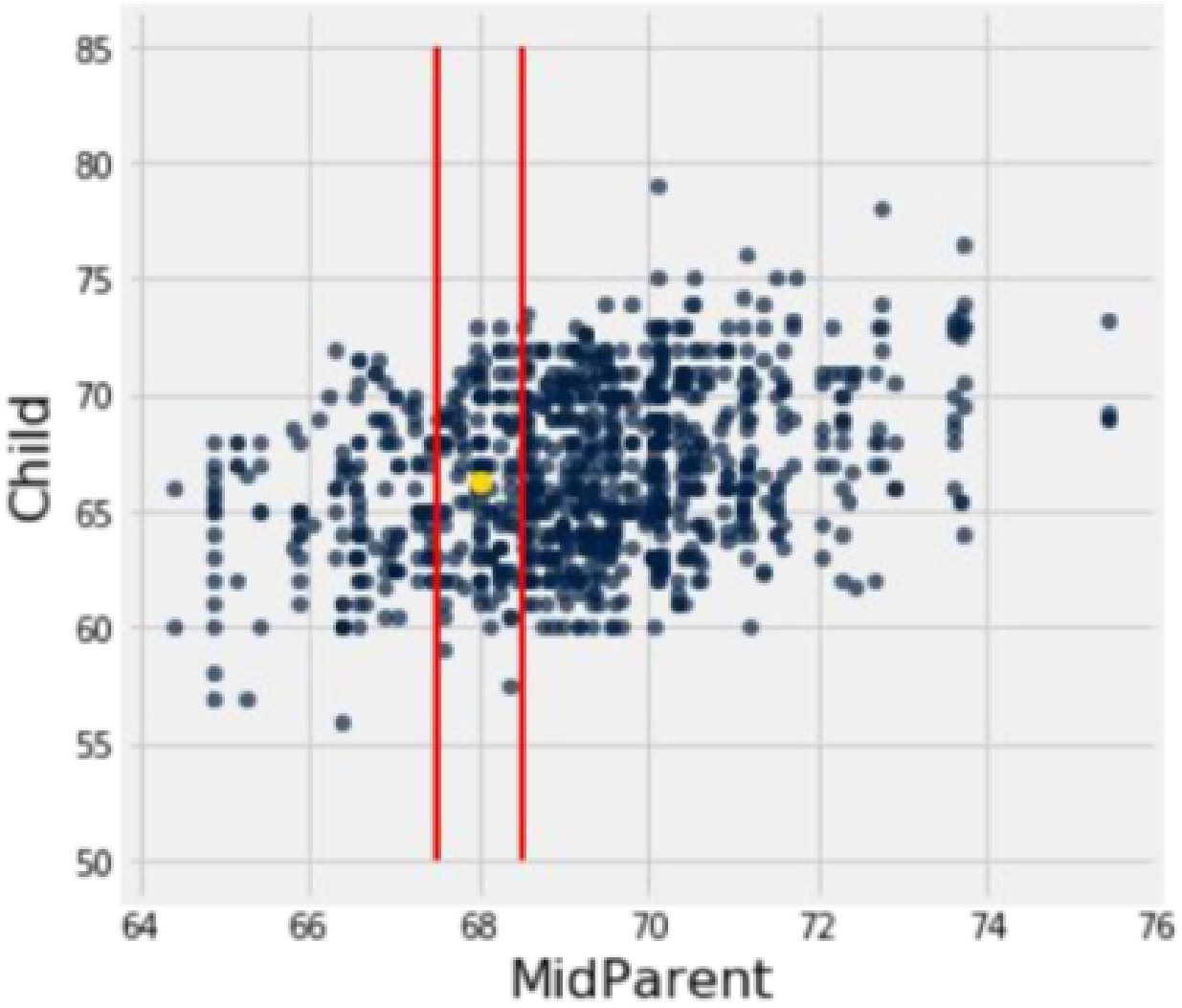
Measures how clustered the scatter is around a straight line

Prediction

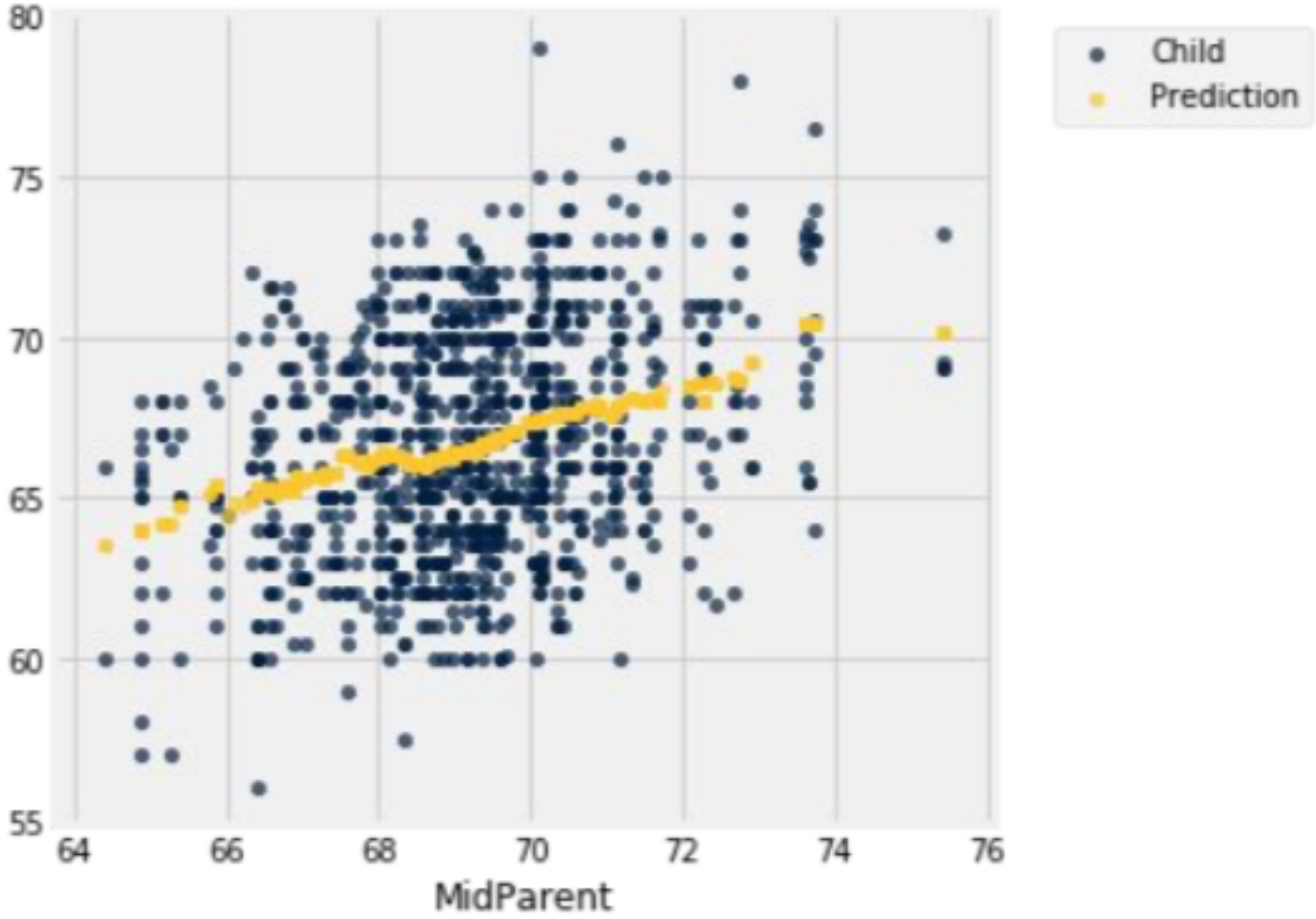
Galton's Heights



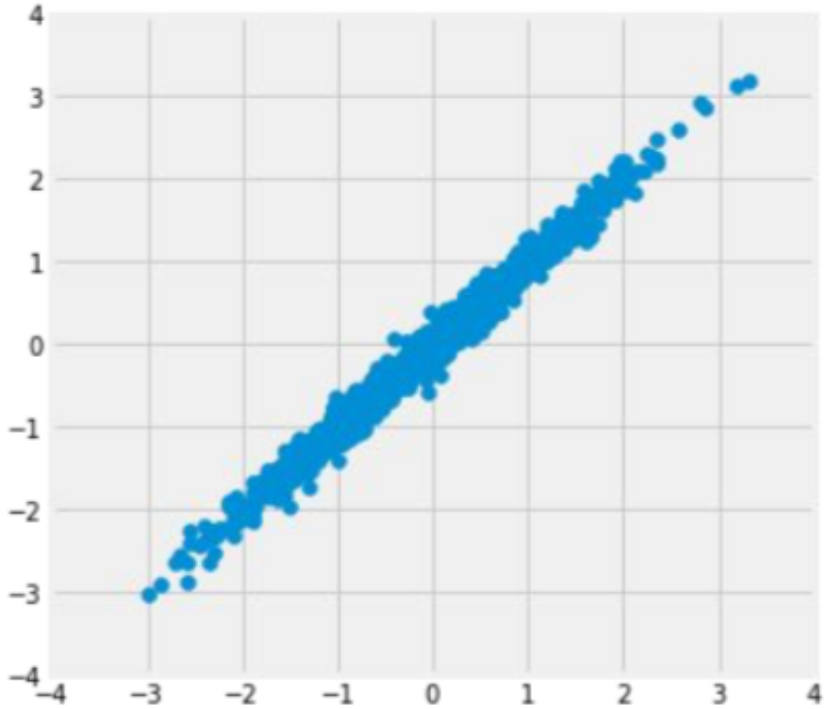
Galton's Heights



Galton's Heights

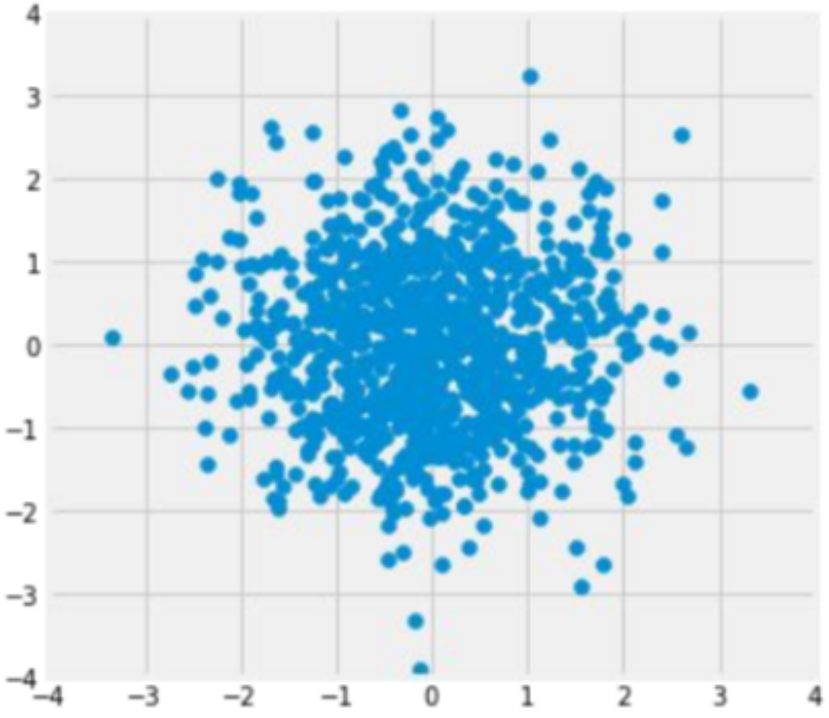


Where is the prediction line?



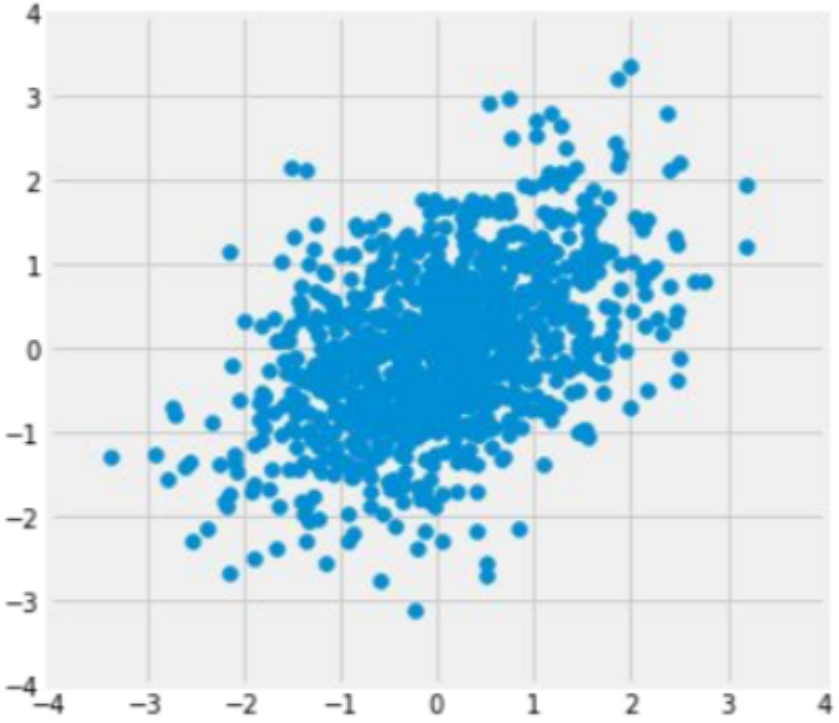
$r = 0.99$

Where is the prediction line?



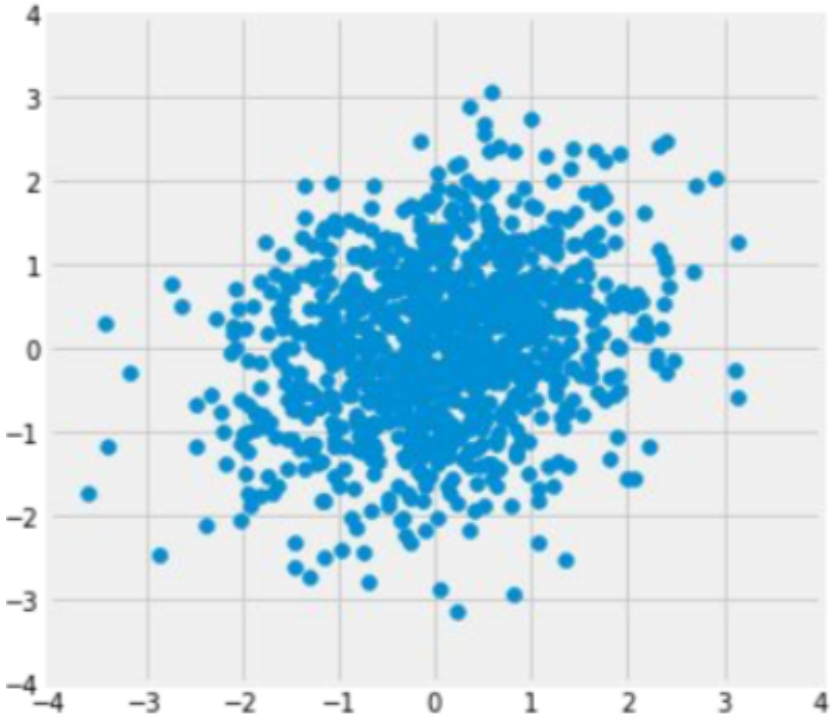
$r = 0.0$

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

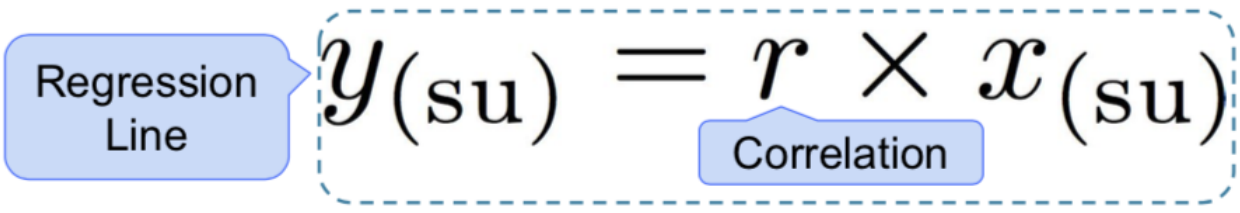
(DEMO)

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



Not true for all points – a statement about averages

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

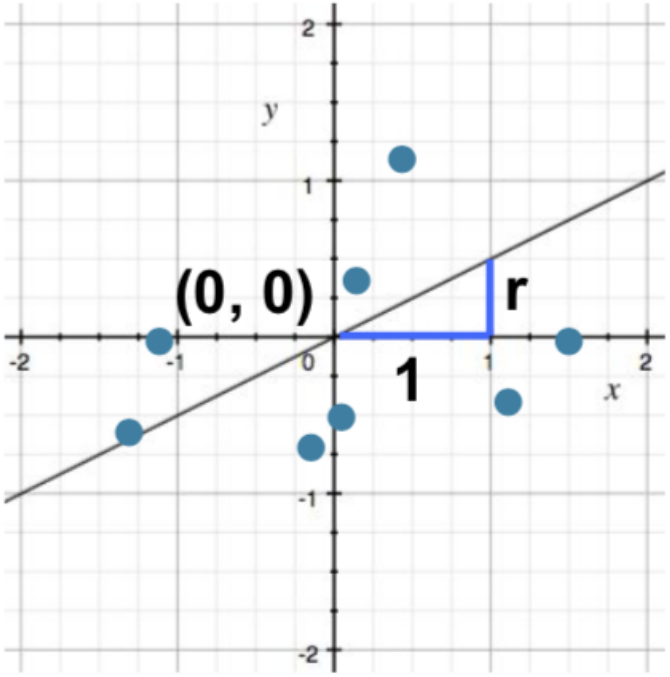
x in standard units

Lines can be expressed by slope & intercept

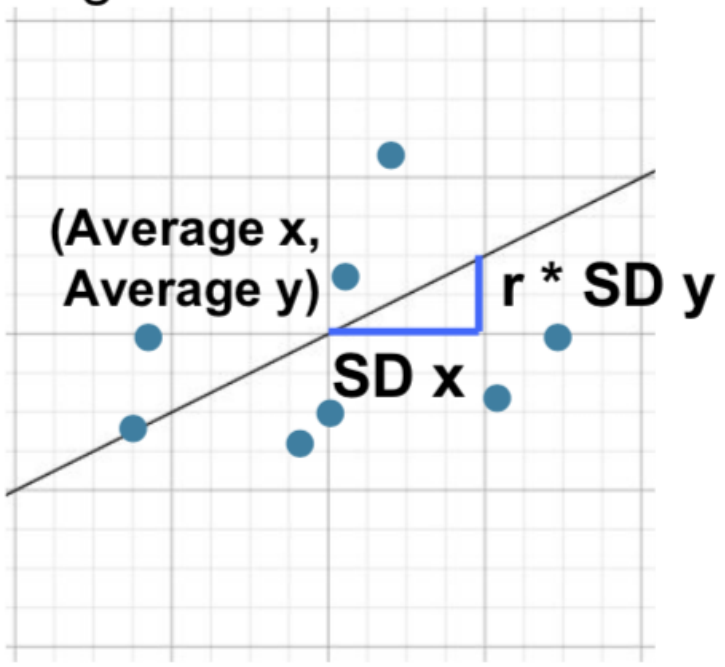
$$y = \text{slope} \times x + \text{intercept}$$

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)