Lecture 28

Linear Regression

Correlation (Review)

The Correlation Coefficient *r*

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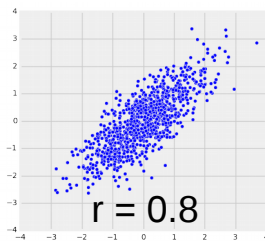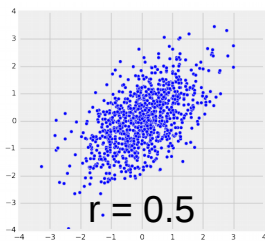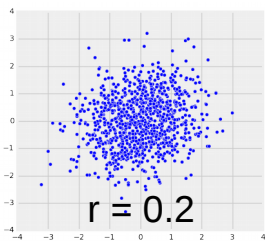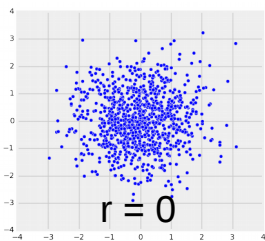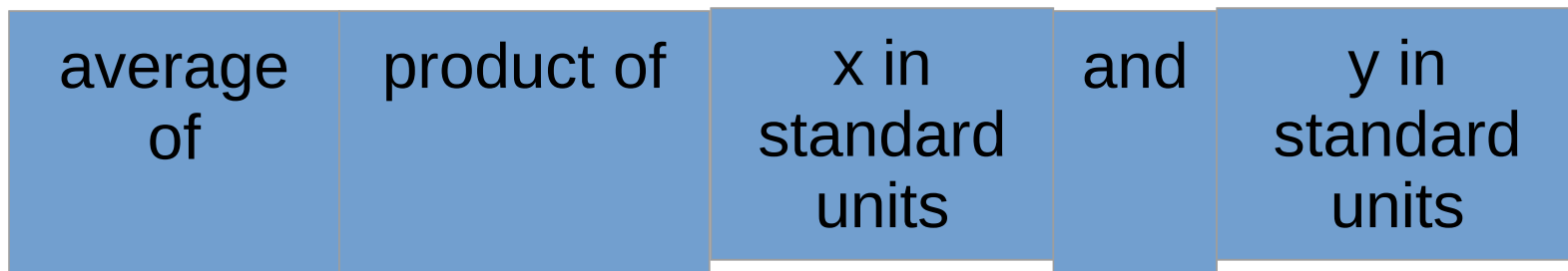- Based on standard units
- -1 ≤ *r* ≤ 1
  - *r* = 1: scatter is perfect straight line sloping up
  - *r* = -1: scatter is perfect straight line sloping down
- *r* = 0: No linear association; *uncorrelated*
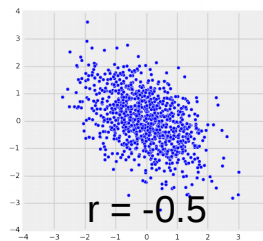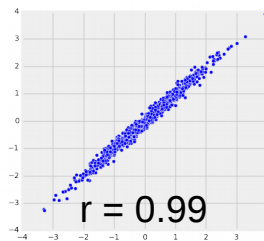
Definition of *r*

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

Properties of Correlation

Properties of *r*

- *r* is a pure number, with no units
- *r* is not affected by changing units of measurement
- *r* is not affected by switching the horizontal and vertical axes

Interpreting *r*

Watch out for:

- Jumping to conclusions about causality
- Non-linearity
- Outliers
- Ecological correlations, based on aggregates or averaged data

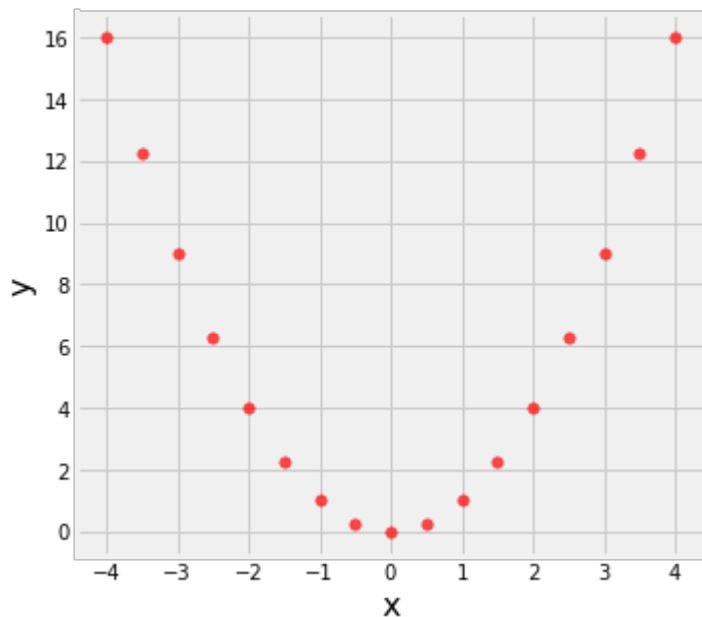# Interpreting *r*

Don't jump to conclusions about causality
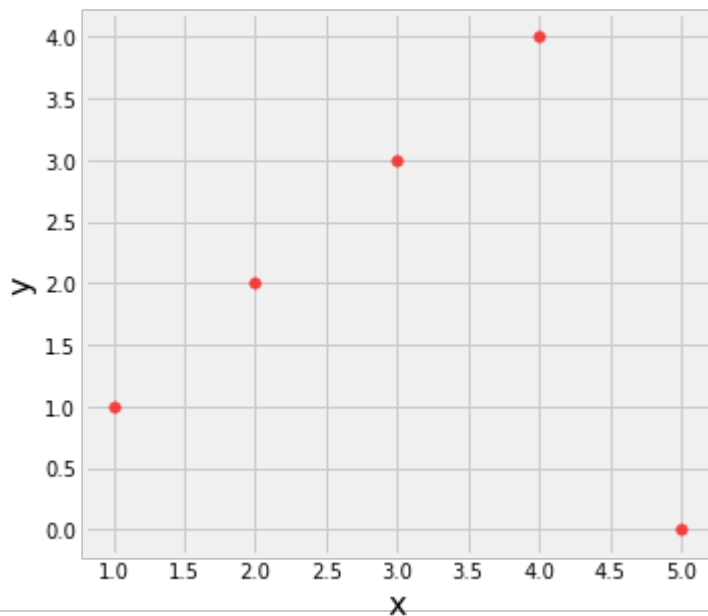
Interpreting *r*

Watch out for non-linearity.

r = 0.0

Interpreting *r*

Watch out for outliers.
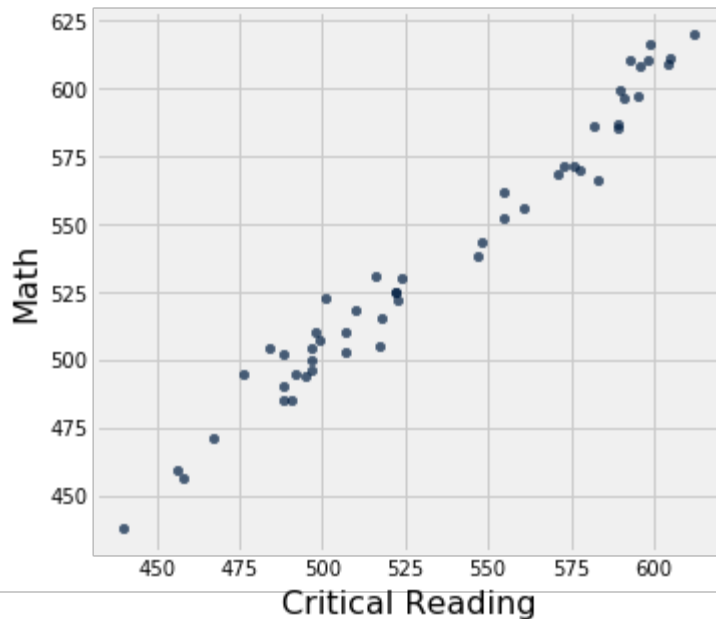

r = 0.0

Interpreting *r*

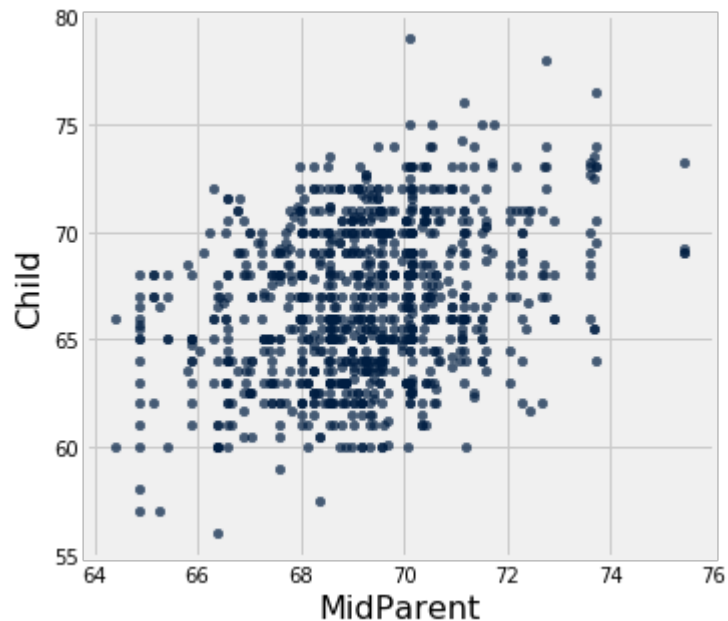Watch out for ecological correlations, based on aggregates or averaged data.

r = 0.98

Prediction

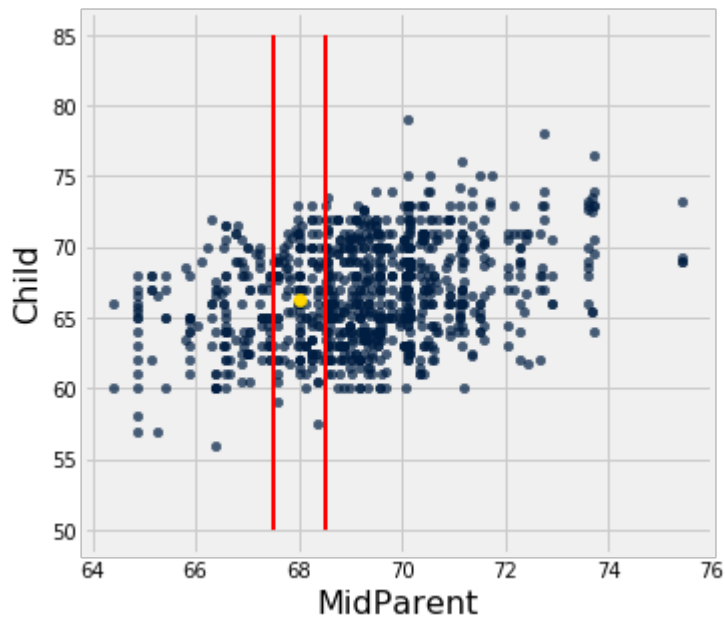# Galton's Heights

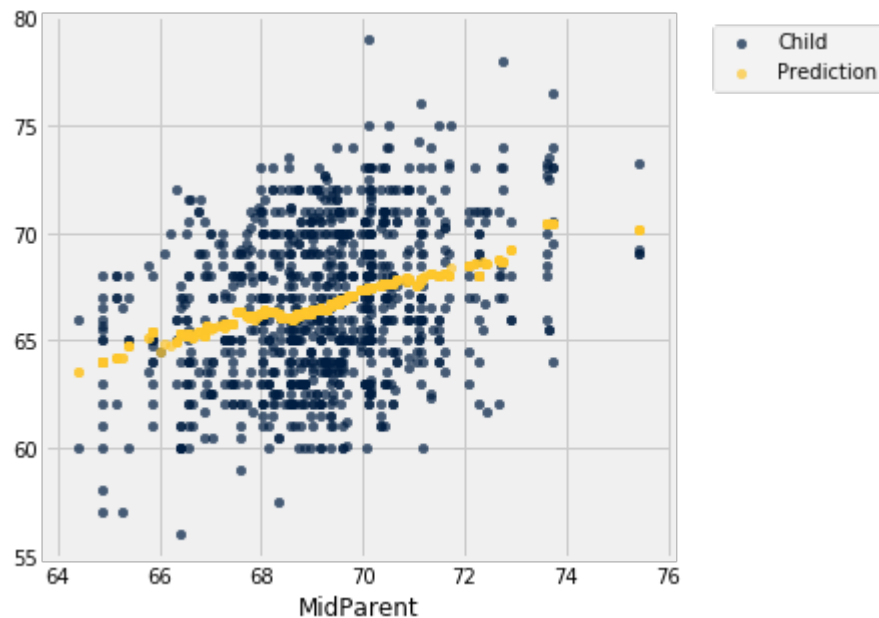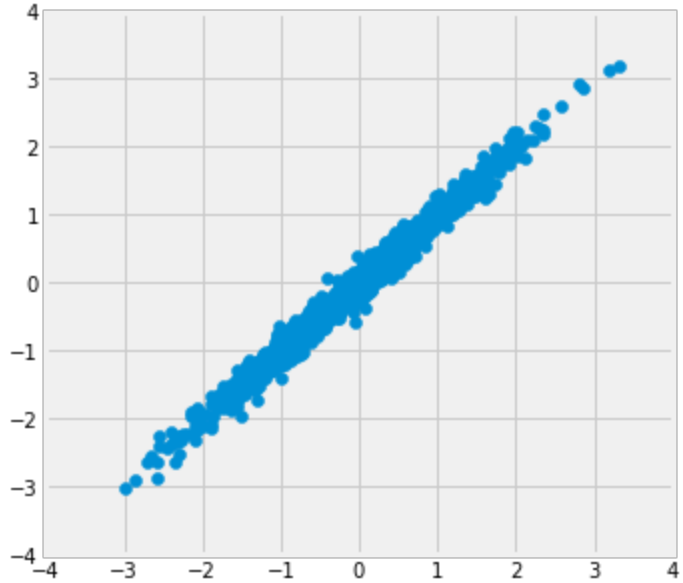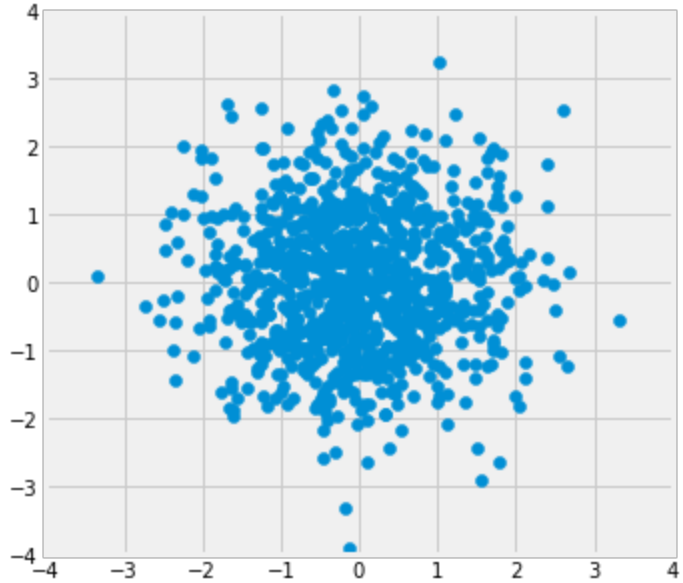# Galton's Heights

Galton's Heights

Where is the prediction line?

r = 0.99

Where is the prediction line?

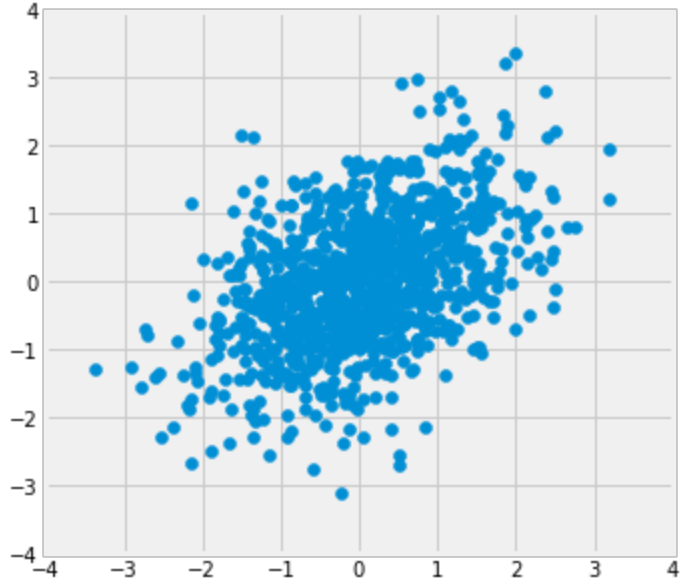

r = 0.0

Where is the prediction line?

r = 0.5

Where is the prediction line?
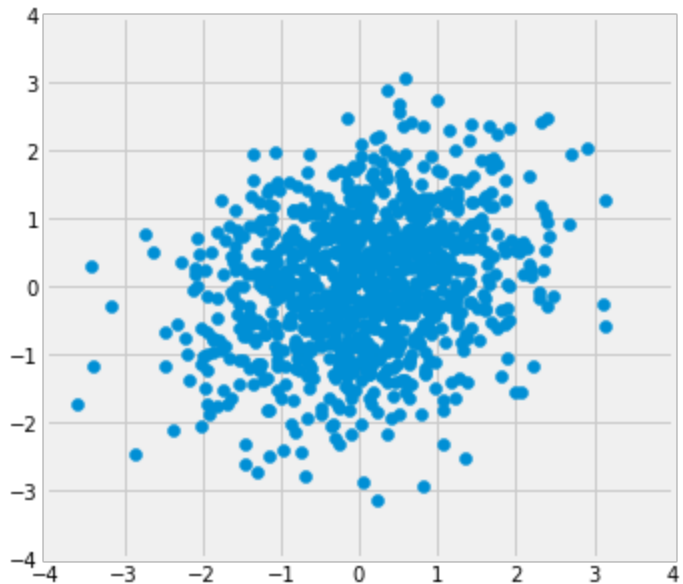
r = 0.2

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.

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

$$y_{(su)} = r \times x_{(su)}$$

Correlation

Not true for all points — a statement about averages

Linear Regression

(Demo)

Slope & Intercept

Regression Line Equation

In original units, the regression line has this equation:

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

y in standard units

x in standard units
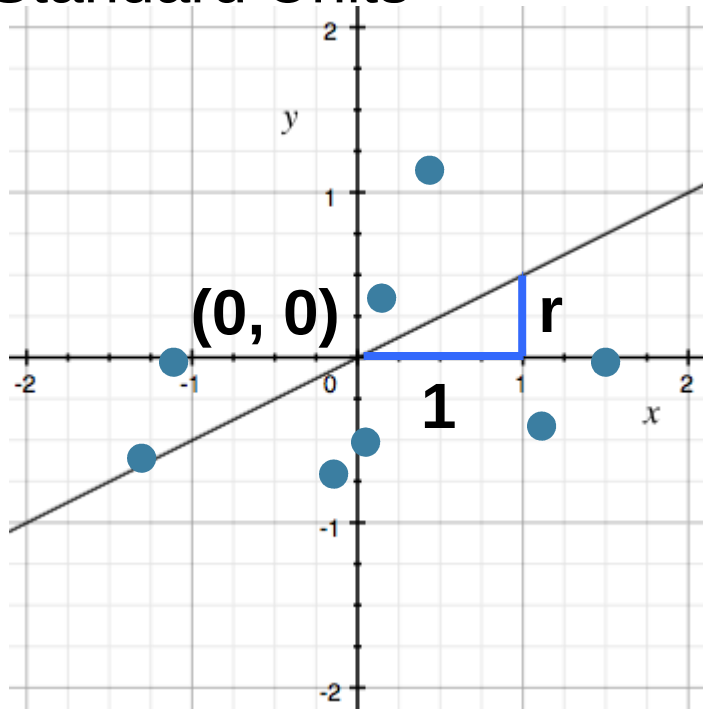
Lines can be expressed by *slope* & *intercept*
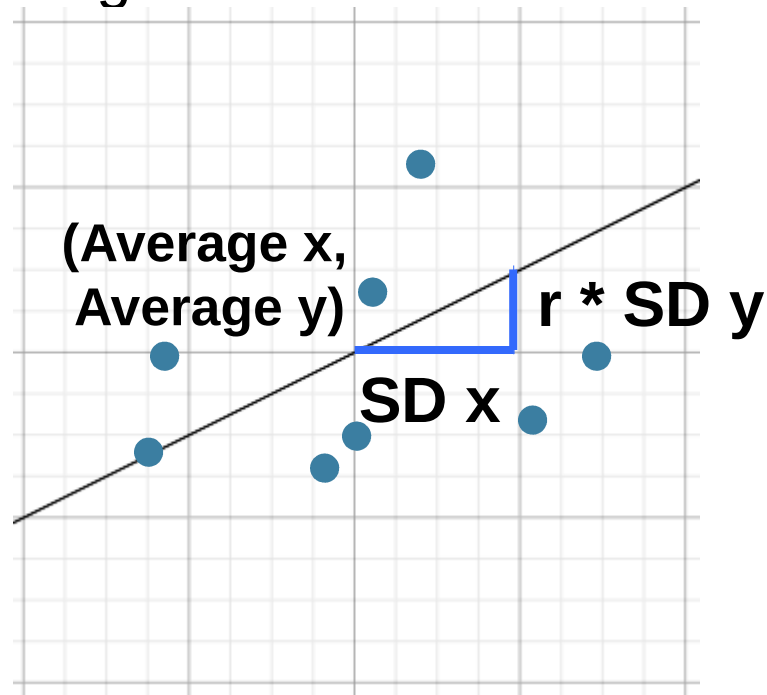
$$y = \text{slope} \times x + \text{intercept}$$

Regression Line

Standard Units

Original Units

Slope and Intercept

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)