Prediction:
Correlation and Linear Regression

Logistics

- Sample Comparison write up due today
- Homework 4 due on Thursday

Prediction

- Predicting one characteristic based on another:
  - Given my height, how tall will I be next year?
  - Given my height, how tall will my kid be as an adult?
  - Given my height, how much will I spend on a boat?

- There's something I know, and something I want to determine
  - Characteristics of an example: known and unknown

- Assumption of prediction: for some sample, we know all the characteristics

Relation Between Two Variables

- Association
- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape”
  - Linear
  - Non-linear
- Good protocol: visualize first, then quantify

(Demo)

The Correlation Coefficient, *r*

- Measures linear association
- Based on standard units
- -1 <= *r* <= 1
  - *r* = 1: scatter plot is perfect straight line sloping up
  - *r* = -1: scatter plot is perfect straight line sloping down
  - *r* = 0: no linear association; *uncorrelated*

(Demo)

The Correlation Coefficient, *r*

1. Convert both variables to standard units
   - Subtract off the mean, divide by the standard deviation
2. Multiply them together
3. Average the products
   - That's *r*

The Correlation Coefficient, *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 the horizontal and vertical axes

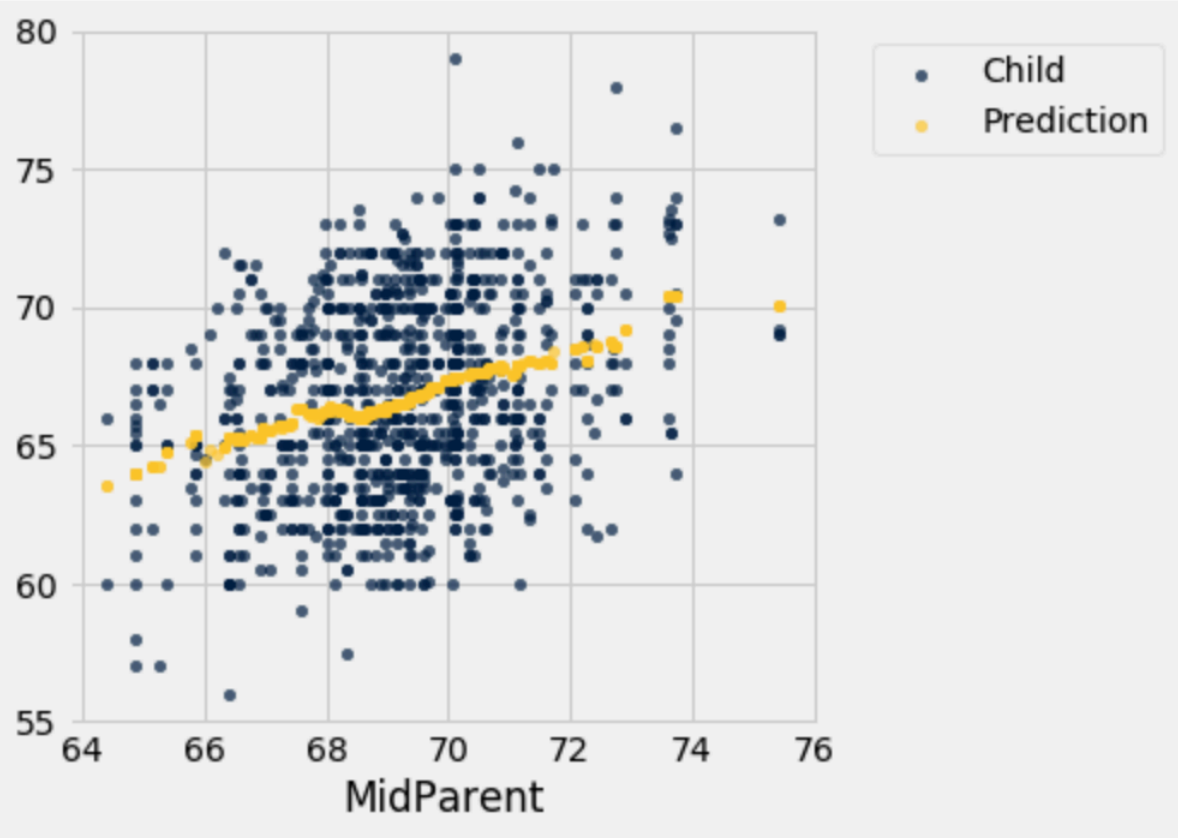
The Correlation Coefficient, *r*

Watch out for:

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

Linear Regression

- Revisit our example from way back when – Galton height prediction



(Demo)

Regression Line Equation

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

Regression Line Equation

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

y in standard units

x in standard units

Regression Line Equation

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

y in standard units

x in standard units

Lines can be expressed by *slope* and *intercept* in original units:

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

Regression Line Equation

Lines can be expressed by *slope* and *intercept* in original units:

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

Slope: $$r \cdot \frac{\text{STD of } y}{\text{STD of } x}$$

Intercept: $$\text{average of y} - \text{slope} \cdot \text{average of } x$$