DATA 8

Fall 2016

Lecture 28, October 31

Correlation

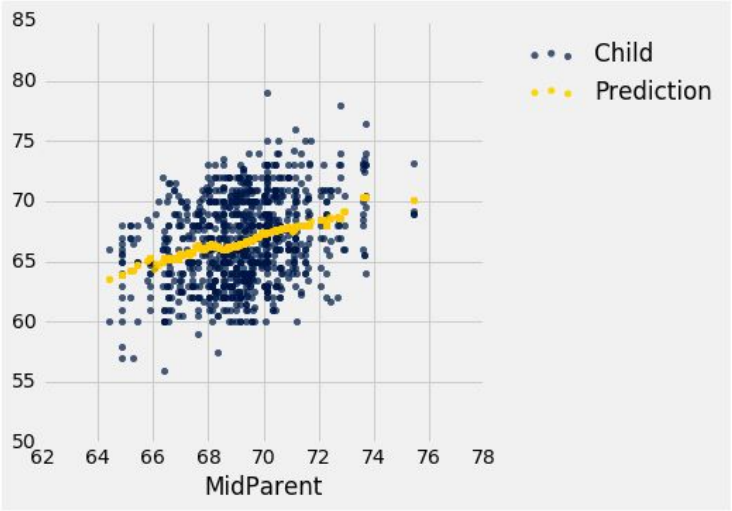
Slides created by Ani Adhikari and John DeNero

Announcements

- Project 2 checkpoint is tomorrow 11/1 at 7 p.m.
- Homework due this week as usual.

Prediction

- Guess outcomes in the future, based on available data
- One simple goal:
  - Predict the value of one variable based on another



Relation between two variables

- Association
- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape”
  - Linear
  - Non-linear

Visualize, then quantify (Demo)

The Correlation Coefficient *r*

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

(Demo)

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

Further 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

(Demo)

Care in the Use of *r*

Watch out for:

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