DSFA

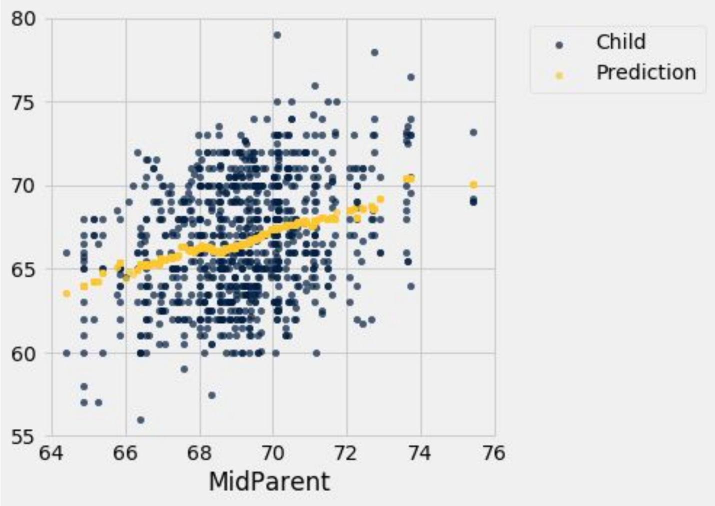
Spring 2020

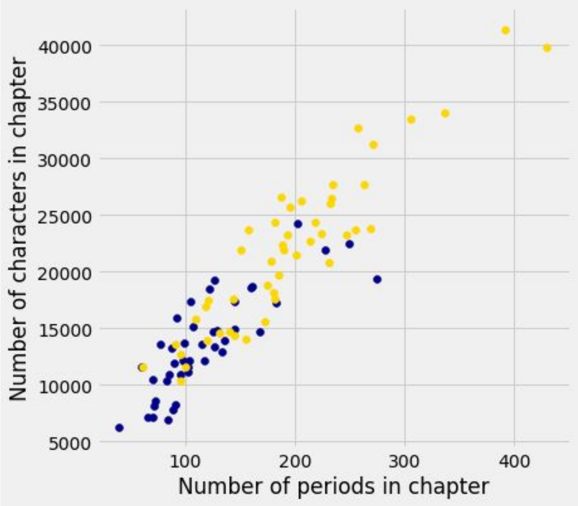
Lecture 22

Correlation

Prediction

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

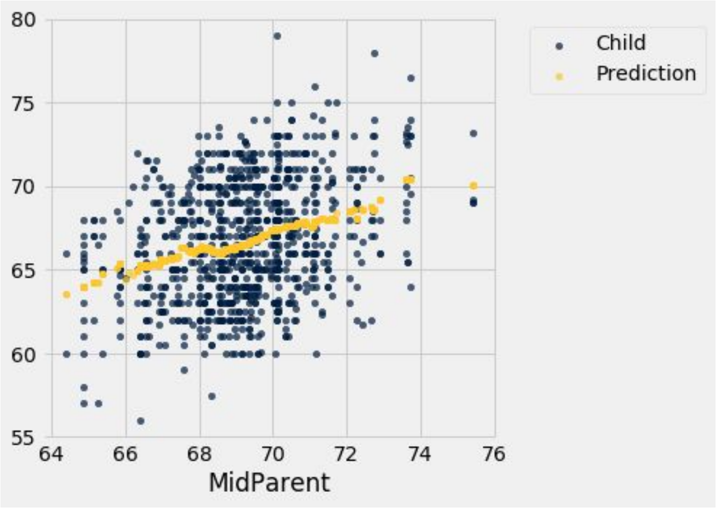


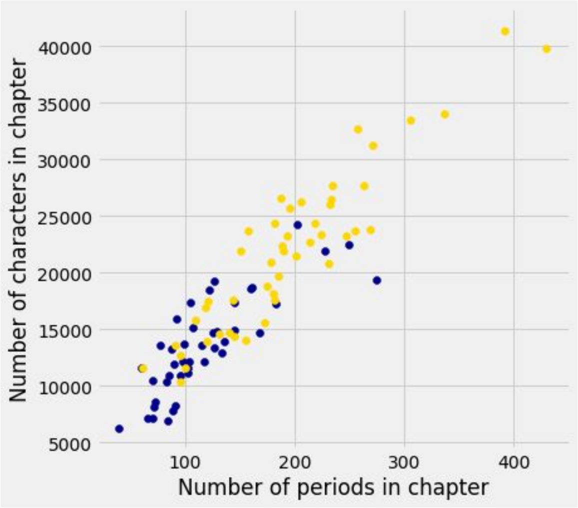


(Demo)

Prediction

If we have a line describing the relation between two variables, we can make predictions





Relation Between Two Variables

Visualize then quantify

- Any discernible pattern?
- Simplest kind of pattern: Linear? Non-linear?

(Demo)

The Correlation Coefficient *r*

- Developed by Karl Pearson (1857-1936) based on work of Francis Galton (1822-1911)
- Measures linear association
- $-1 \leq r \leq 1$
  - $r = 1$: scatter is perfect straight line sloping up
  - $r = -1$: scatter is perfect straight line sloping down
- $r = 0$: No linear association; *uncorrelated*

(Demo)

Definition of *r*

Correlation Coefficient (*r*) =

average of	(array) product of	x in standard units	and	y in standard units

Measures how clustered the scatter is around a straight line

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)

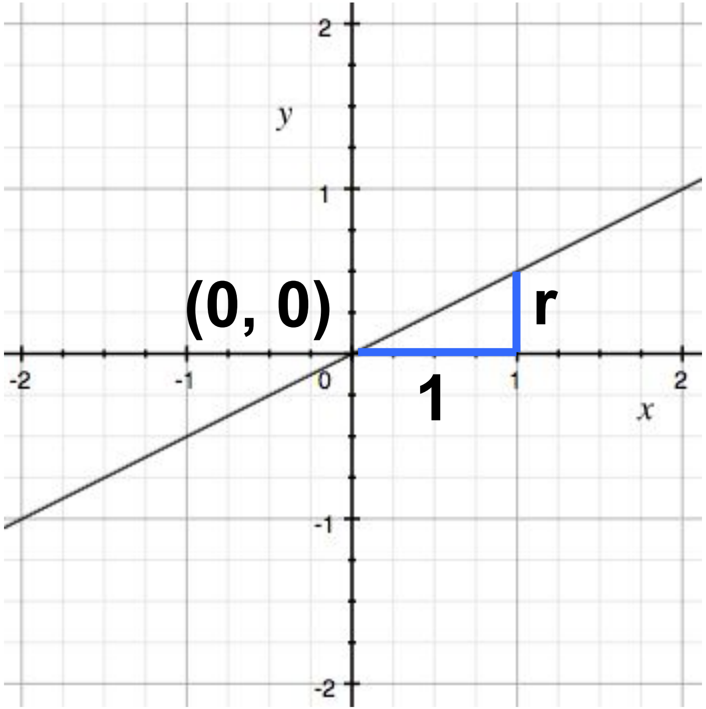
Prediction

Prediction

- **Problem:** given a known x value, predict y, where both are in standard units
- **Solution:**
  - Compute r
  - Predict that $y = r * x$
- Why is that a line?

Algebra review:

Equation of a Line



$$y = r * x$$

In general:

$$y = a * x + b$$

(a is slope, b is intercept)

Prediction

- **Problem:** given a known x value, predict y, where both are in standard units
- **Solution:**
  - Compute r
  - Predict that $y = r * x$
- Why is that a line?
- Why use *that* line?

(Demo)

Prediction

- **Problem:** given a known x value, predict y, where both are in standard units
- **Solution:**
  - Compute r
  - Predict that $y = r * x$
- Why is that a line?
- Why use *that* line?
  - It is a version of the graph of averages, smoothed to a line

(Demo)

Prediction

- **Predict** $y = r * x$ (in standard units)

- Example:
  - $x = 2$ (in standard units)
  - $r = .75$
  - What is the prediction for y (in standard units)?
    - A. 0.0
    - B. 0.75
    - C. 1.5
    - D. 2.0

Prediction

- **Predict** $y = r * x$ (in standard units)

- Example:
  - A course has a typical prelim (mean=70, std=10), and a hard final (mean=50, std=12)
  - The scores on the exams look linearly related when visualized, with $r = .75$
  - **Predict** a student's final exam score, given that their prelim score was 90 *(go ahead and work on that)*

Prediction

- Prelim: mean=70, std=10
  - x = 90 = 70 + 2*10 in original units = 2 standard units
- Prediction:
  - y = r * x = .75 * 2 = 1.5 standard units
- Final: mean=50, std=12
  - y = 50 + 1.5 * 12 = **68** in original units

Prediction

- **Predict** $y = r * x$ (in standard units)
- If r = .75 and x is 2 std above mean,
  then prediction for y is 1.5 std above mean
- So y predicted to be closer to mean than x

- "Regression to the mean"
  - Children with exceptionally tall parents tend not to be as tall
  - Galton called it "regression to mediocrity"

(Demo)