DSC 10, Spring 2018
Lecture 21

Correlation

sites.google.com/eng.ucsd.edu/dsc-10-spring-2018

Credit: Anindita Adhikari and John DeNero

Review:
Choosing a Sample Size

Designing your sample

- You want to estimate what proportion of voters will vote for Candidate A in an upcoming election.

- How many people should you sample at random in order to get a 95% confidence interval with a width of 0.03 or less?

Control the Width of 95% CI

- Suppose you want a width of no more than 0.03
- 95% CI is Mean **± 2 SDs** of the sample proportion
- Total width = 4 SDs of the sample proportion

  $$= 4 \times (\text{population SD})/\sqrt{(\text{sample size})} \leq 0.03$$

- Solve for sample size

  $$\text{sample size} \geq (4 \times (\text{population SD}) / 0.03)^2$$

Bound the Population SD

Problem: We don't know the population SD.

Fact: SD of population of 0's and 1's is always ≤ 0.5

$$\text{sample size} \geq (4 \text{ x } (0.5) \;/\; 0.03)^2$$

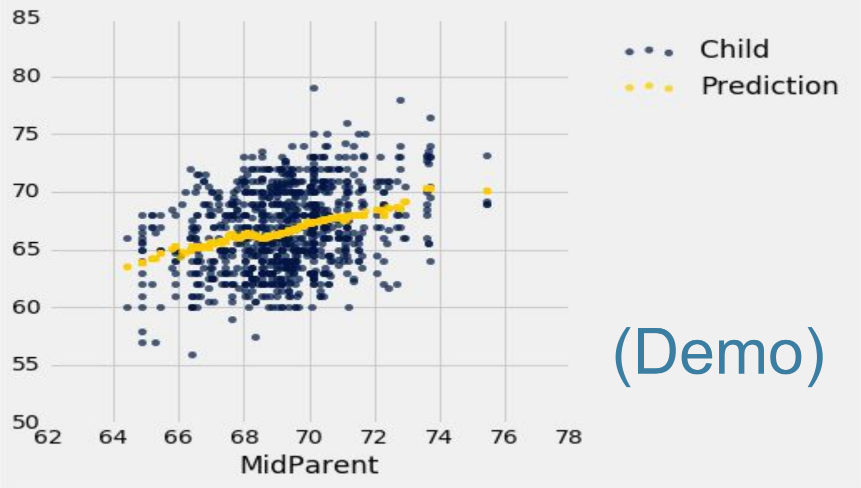
$$\geq (4 \text{ x } (\text{population SD}) \;/\; 0.03)^2$$

Choose a sample size of at least $(4 \text{ x } (0.5) / 0.03)^2 = 4445$.

Prediction

Prediction Problems

- Predicting one characteristic based on another:
  - Given my height, how tall will my kid be as an adult?
  - Given my education level, what is my income?
  - Given my income, how much does my car cost?
- Two characteristics: one is known, one is unknown
- Have some data for which we know both characteristics
- To predict, need an association



(Demo)

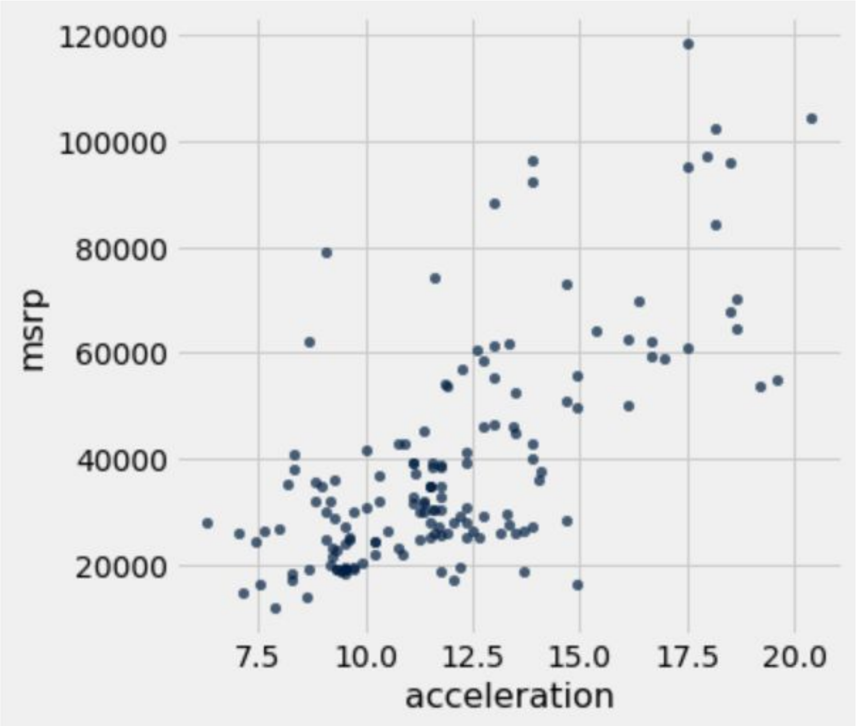
Correlation

Relation Between Two Variables

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

Visualize then quantify

Association



This scatter plot shows that people are generally

A. Willing to pay more for cars that accelerate faster

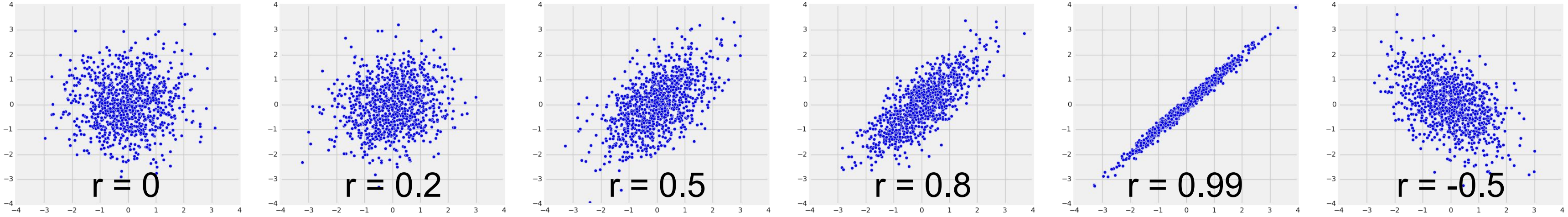
B. Willing to pay more for certain cars because they accelerate faster

C. Not willing to pay more for cars that accelerate faster

D. More than one of the above

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

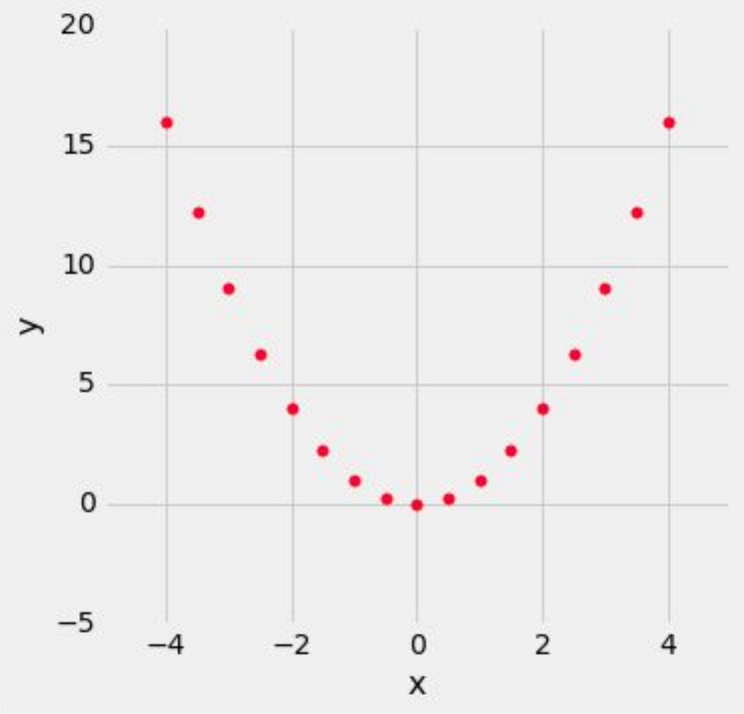
Measures how clustered the scatter is around a straight line

r: average of product of standard units

x	y	x (standard units)	y (standard units)	product of standard units
1	2	-1.46385	-0.648886	0.949871
2	3	-0.87831	-0.162221	0.142481
3	1	-0.29277	-1.13555	0.332455
4	5	0.29277	0.811107	0.237468
5	2	0.87831	-0.648886	-0.569923
6	7	1.46385	1.78444	2.61215

- Then calculate the average = 0.617

Question: $y = x^2$



This scatter plot shows

A. association and correlation
B. association but not correlation
C. correlation but not association
D. neither association nor correlation

Linear Regression

Graph of Averages

A visualization of x and y pairs

- Group each x value with other nearby x values
- Average the corresponding y values for each group
- For each x value, produce one predicted y value

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

Regression Line

$$\boxed{y_{(\text{su})} = \underbrace{r}_{\text{Correlation}} \times x_{(\text{su})}}$$

- If r = 0.6, and the given x is 2 standard units, then:
  - The given x is 2 SDs above average
  - The prediction for y is 1.2 SDs above average

- On average (though not for each individual), regression predicts y to be closer to the mean than x is

Discussion Question

A course has a midterm (average 70; standard deviation 10) and a really hard final (average 50; standard deviation 12)

If the scatter diagram of midterm & final scores for students has a typical oval shape with **correlation 0.75**, then what is the average final exam score for students who scored **90 on the midterm**?

A. 76

B. 90

C. 68

D. 82

E. 67.5

Discussion Question: Solution

A course has a midterm (average 70; standard deviation 10)
and a really hard final (average 50; standard deviation 12)

If the scatter diagram of midterm & final scores for students has a typical oval shape with **correlation 0.75**, then what is the average final exam score for students who scored **90 on the midterm**?

1. (90 - 70)/10 = 2 standard units on midterm,
2. estimate 0.75 * 2 = 1.5 standard units on final
3. estimated final score = 1.5 * 12 + 50 = 68 points

(Demo)

Slope & Intercept

Regression Line Equation

In original units, the regression line has this equation:

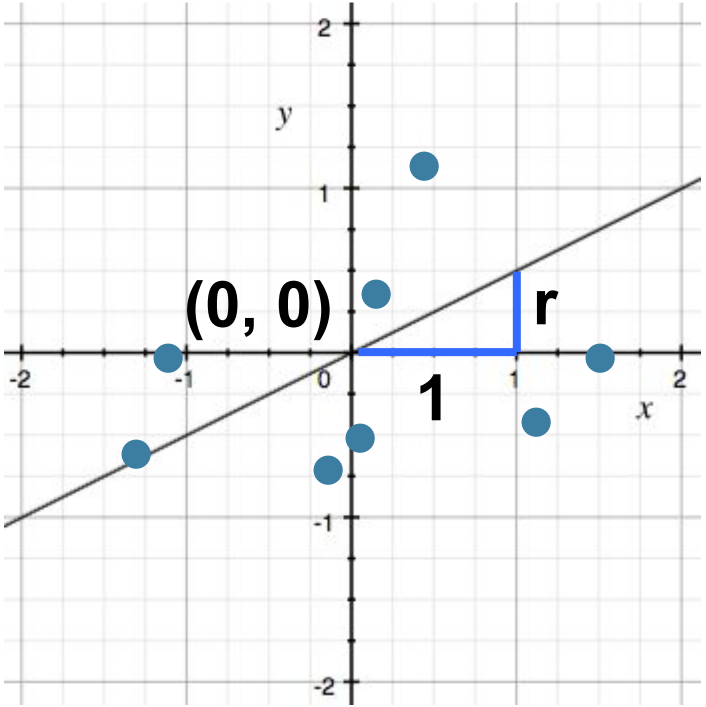
$$\underbrace{\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y}}_{\text{y in standard units}} = r \times \underbrace{\frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}}_{\text{x in standard units}}$$

Lines can be expressed by *slope* & *intercept*

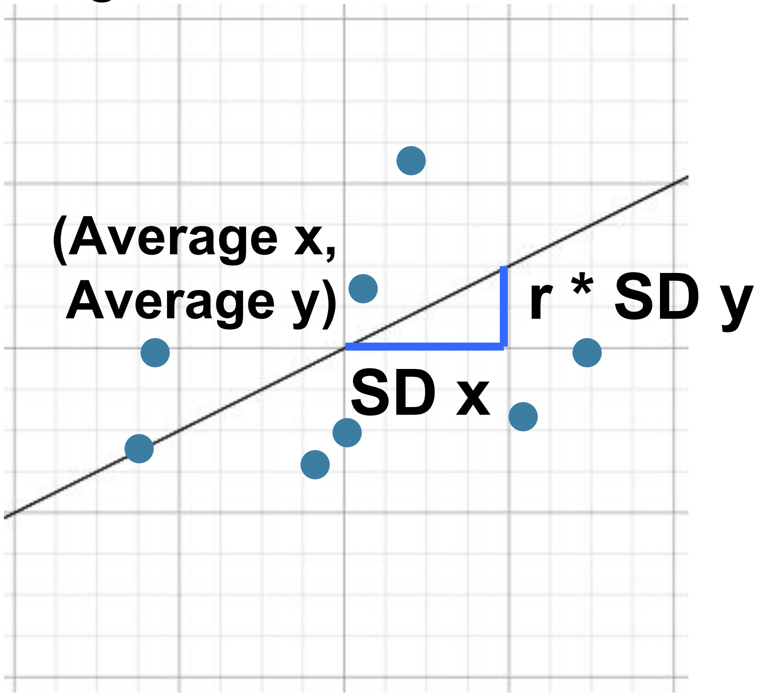
$$y = \text{slope} \times x + \text{intercept}$$

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)