DATA 8

Spring 2020

Lecture 30

Linear Regression

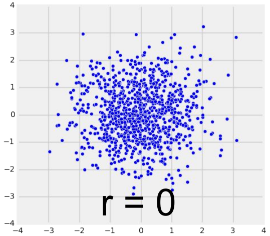
Regression roadmap

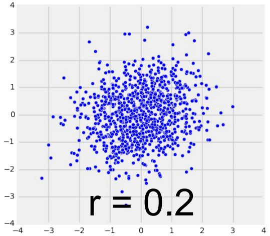
- Monday:
  - Association and correlation
- **Today:**
  - **Prediction, scatterplots and lines**
- Friday:
  - Least squares: finding the “best” line for a dataset
- Next Monday:
  - Residuals: analyzing mistakes and errors
- Next Wednesday:
  - Regression inference: understanding uncertainty

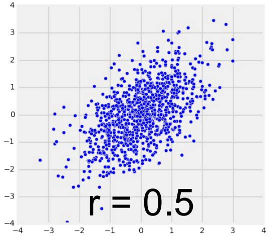
Correlation (Review)

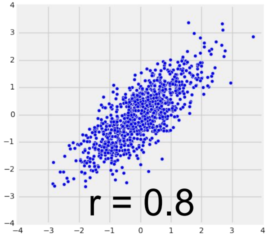
The Correlation Coefficient *r*

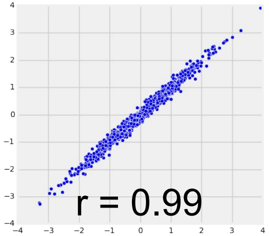
- Measures ***linear*** association
- Based on standard units
- -1 ≤ *r* ≤ 1
  - *r* = 1: scatter is perfect straight line sloping up
  - *r* = -1: scatter is perfect straight line sloping down
- *r* = 0: No linear association; *uncorrelated*

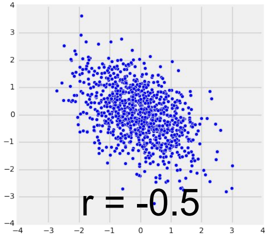












Definition of *r*

Correlation Coefficient (*r*) =

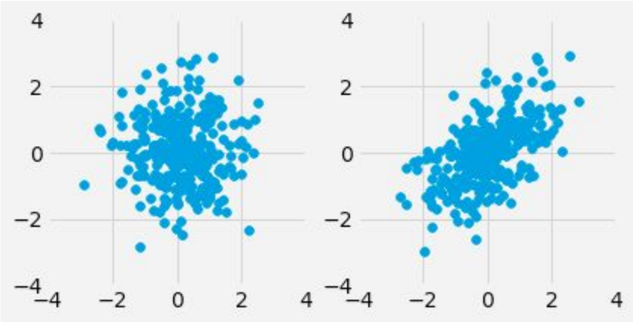
average of	product of	x in standard units	and	y in standard units

Measures how clustered the scatter is around a straight line

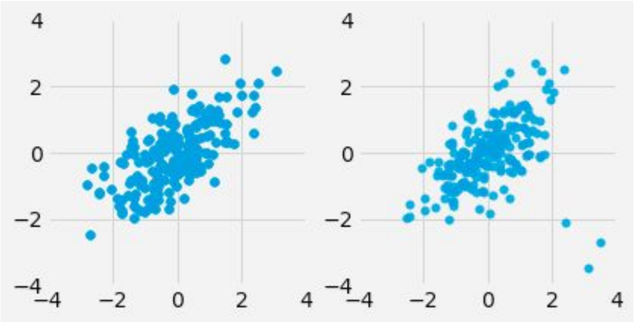
Discussion Question

For each pair, which one will have a higher* value of r?

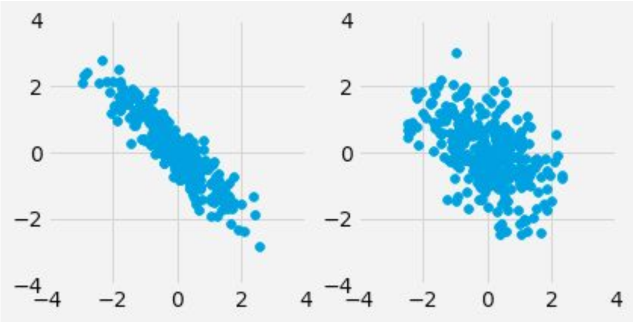
a)



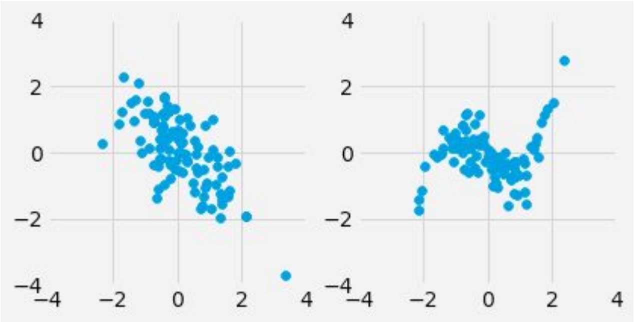
b)



c)



d)



* here, “higher” means “bigger on the number line”

Care in Interpretation

Watch Out For ...

- False conclusions of causation
- Nonlinearity
- Outliers
- Ecological Correlations

(Demo)

Chocolate and Nobel Prizes

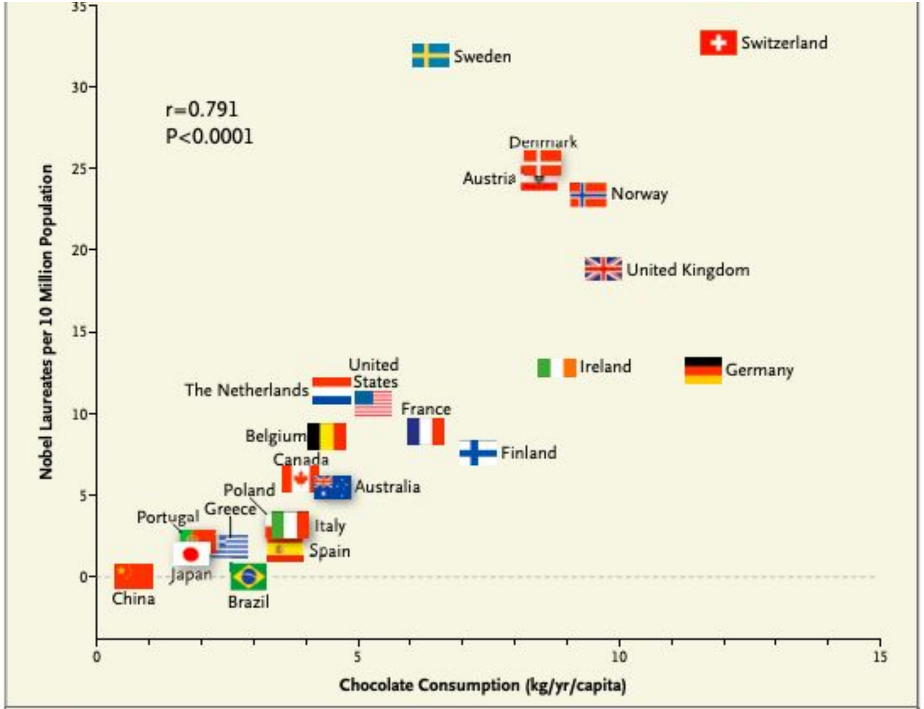


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

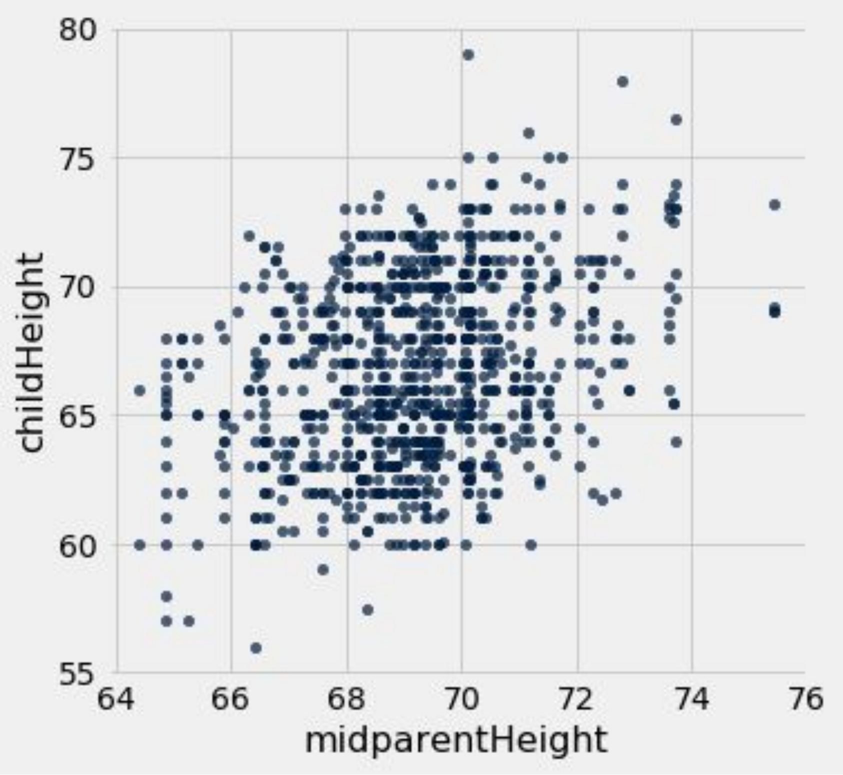
Discussion question

True or False?

1. If x and y have a correlation of 1, then one must cause the other.
2. If the correlation of x and y is close to 0, then knowing one will never help us predict the other.
3. If x and y have a correlation of -0.8, then they have a negative association.

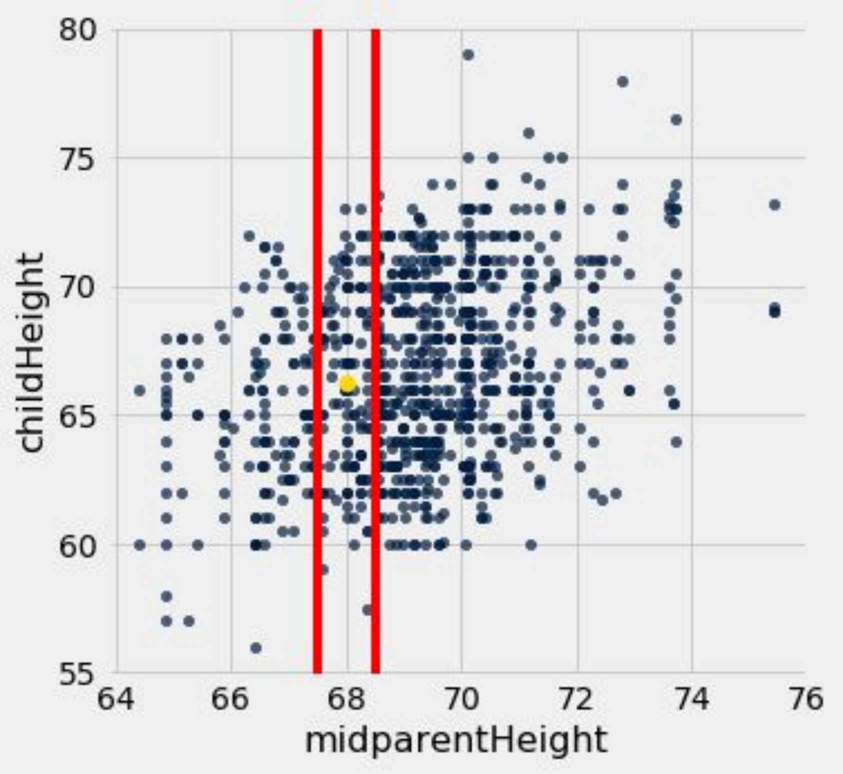
Prediction

Galton's Heights

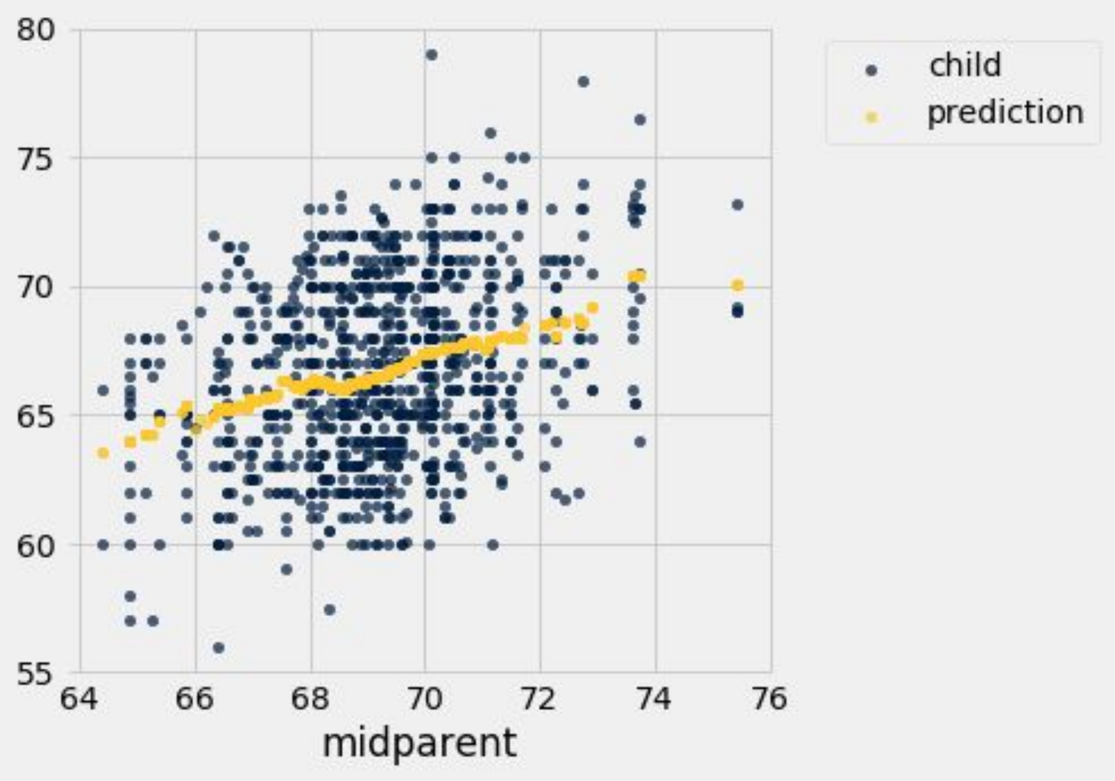


- Oval shaped
- Moderate positive correlation
- How can we predict child height from mid-parent height?

Galton's Heights



Galton's Heights



Nearest Neighbor Regression

A method for prediction:

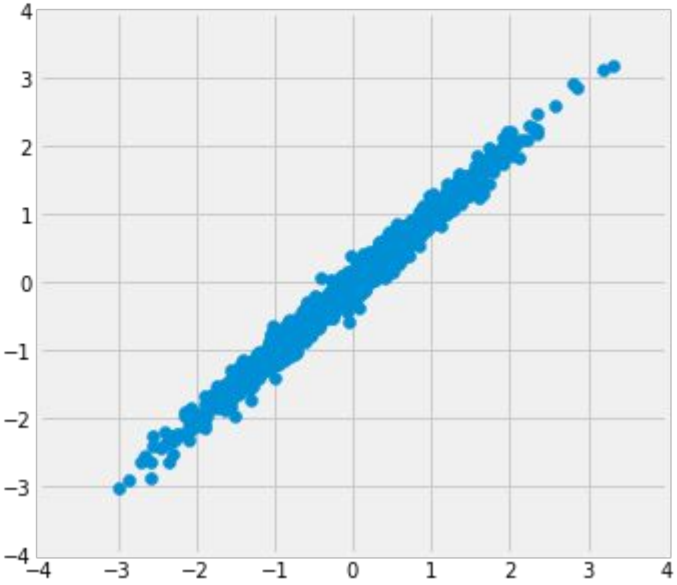
- Group each x with similar (nearby) x values
- Average the corresponding y values for each group

For each x value, the prediction is the average of the y values in its nearby group.

The graph of these predictions is the “graph of averages”.

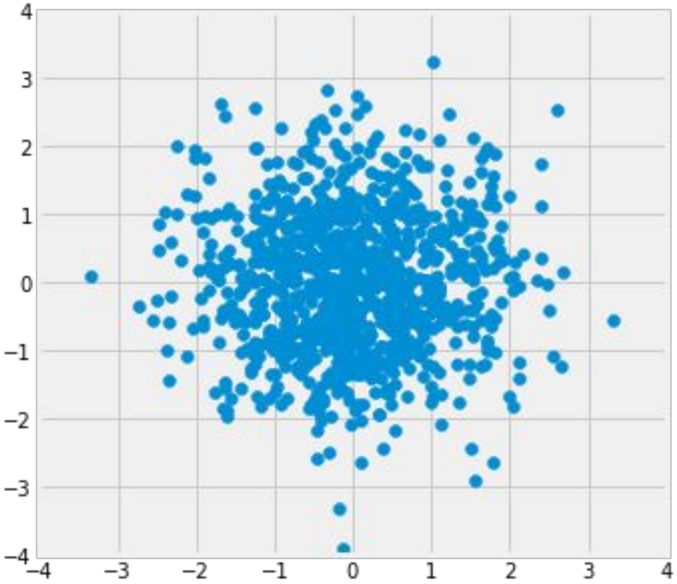
If the association between x and y is linear, then points in the graph of averages tend to fall on a line.

Where is the prediction line?



r = 0.99

Where is the prediction line?



r = 0.0

(Demo)

Linear Regression

Linear Regression

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

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

Not true for all points — a statement about averages

Slope & Intercept

Regression Line Equation

In original units, the regression line has this equation:

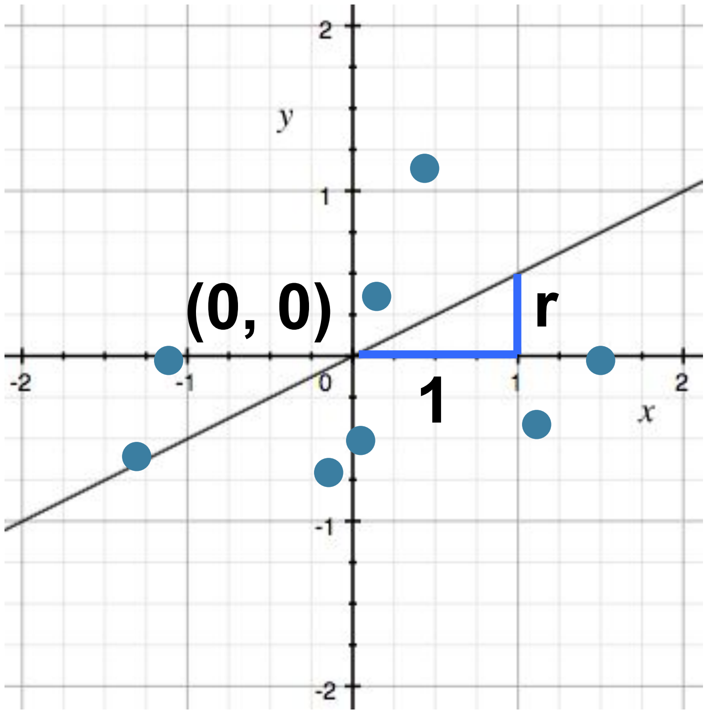
$$\underbrace{\frac{\text{estimate of } y \;-\; \text{average of } y}{\text{SD of } y}}_{\text{estimated y in standard units}} = r \times \underbrace{\frac{\text{the given } x \;-\; \text{average of } x}{\text{SD of } x}}_{\text{x in standard units}}$$

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

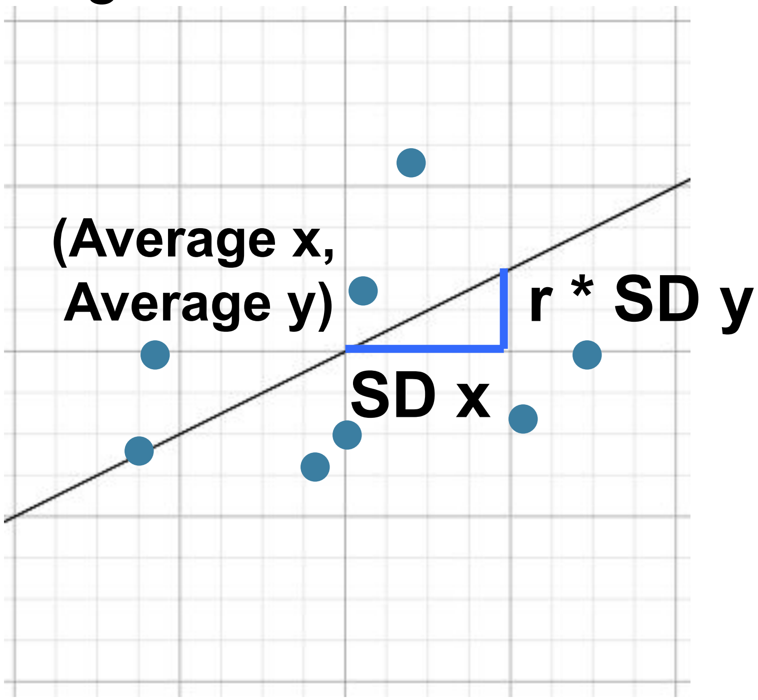
$$y = \text{slope} \times x + \text{intercept}$$

Regression Line

Standard Units

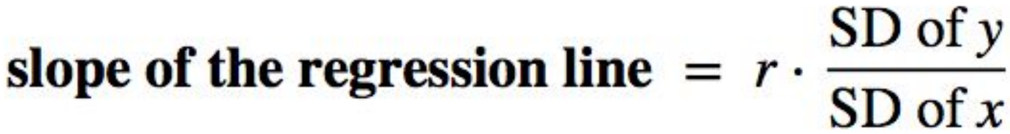


Original Units

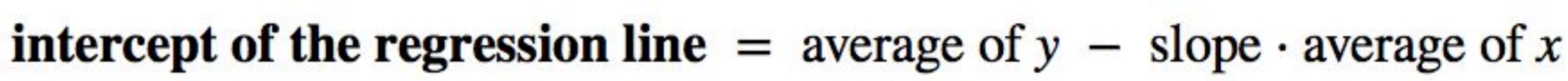


Slope and Intercept

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)

Discussion Question

Suppose we use linear regression to predict candy prices (in dollars) from sugar content (in grams). What are the units of each of the following?

- r
- The slope
- The intercept

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 comparing midterm & final scores for students has an oval shape with correlation 0.75, then...

What do you expect the average final score would be for students who scored 90 on the midterm?

How about 60 on the midterm?

(Demo)