DATA 8

Spring 2022

Lecture 30

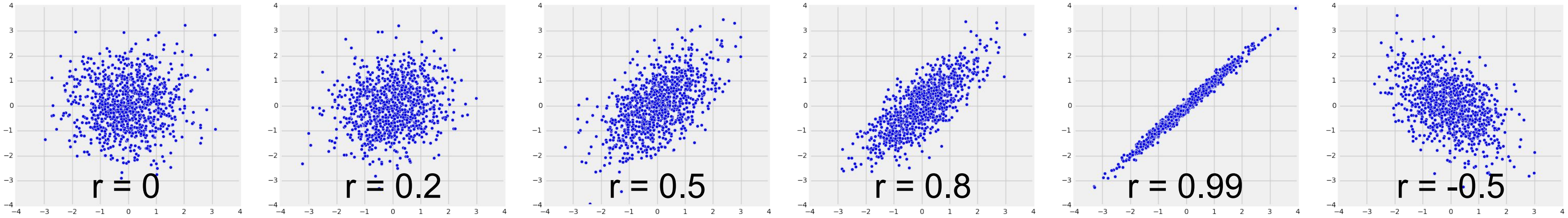
Linear Regression

Announcements

Correlation Coefficient

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

Care in Interpretation

Watch Out For ...

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

(Demo)

Discussion question

True or False?

If the correlation of *x* and *y* is close to 0, then knowing one cannot help us predict the other.

Chocolate and Nobel Prizes

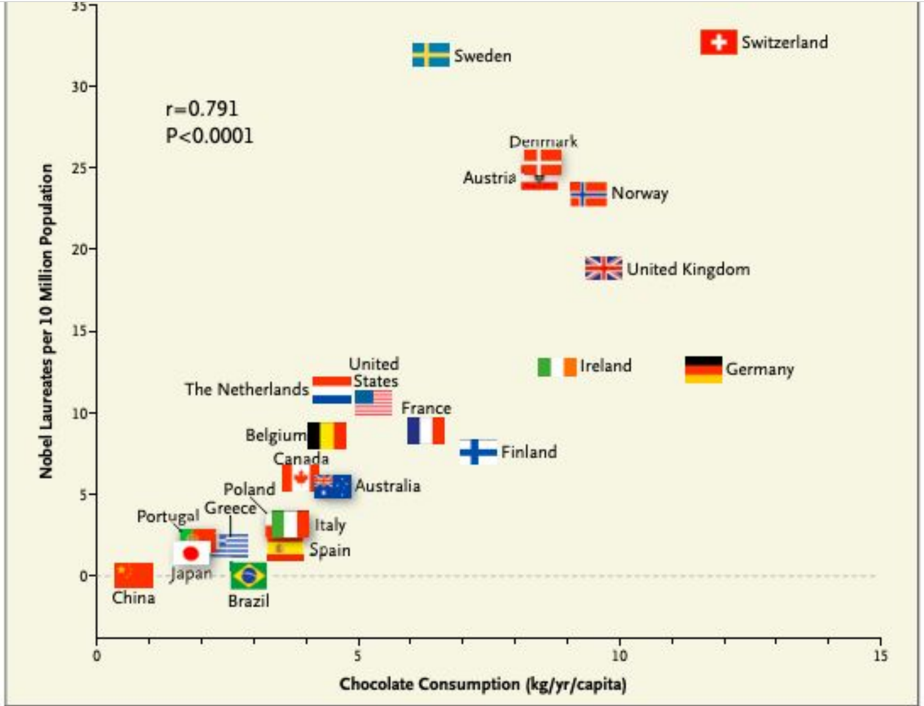
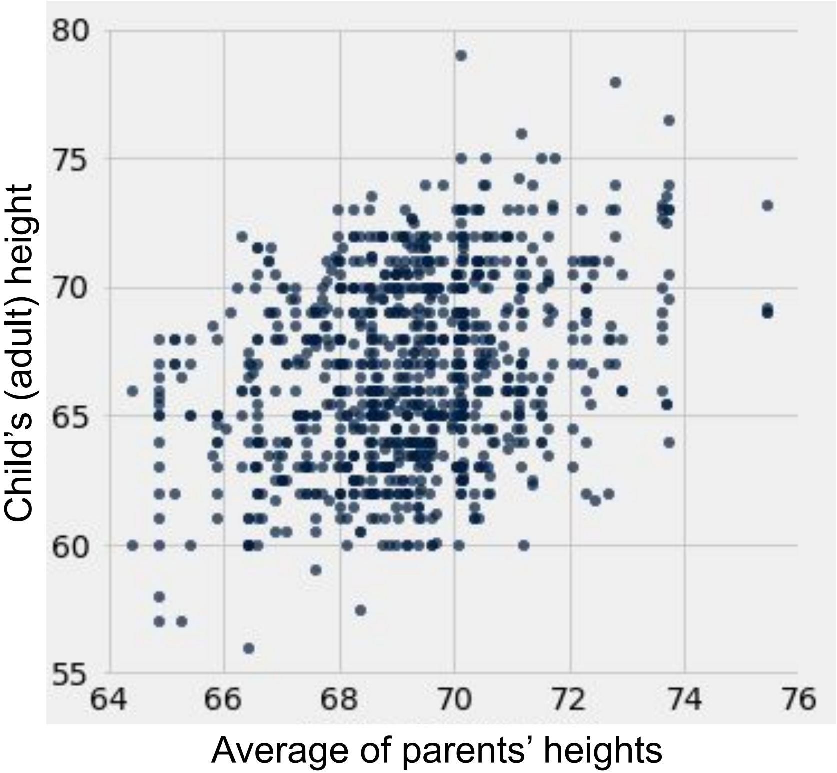


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

https://www.biostat.jhsph.edu/courses/bio621/misc/Chocolate%20consumption%20cognitive%20function%20and%20nobel%20laurates%20(NEJM).pdf

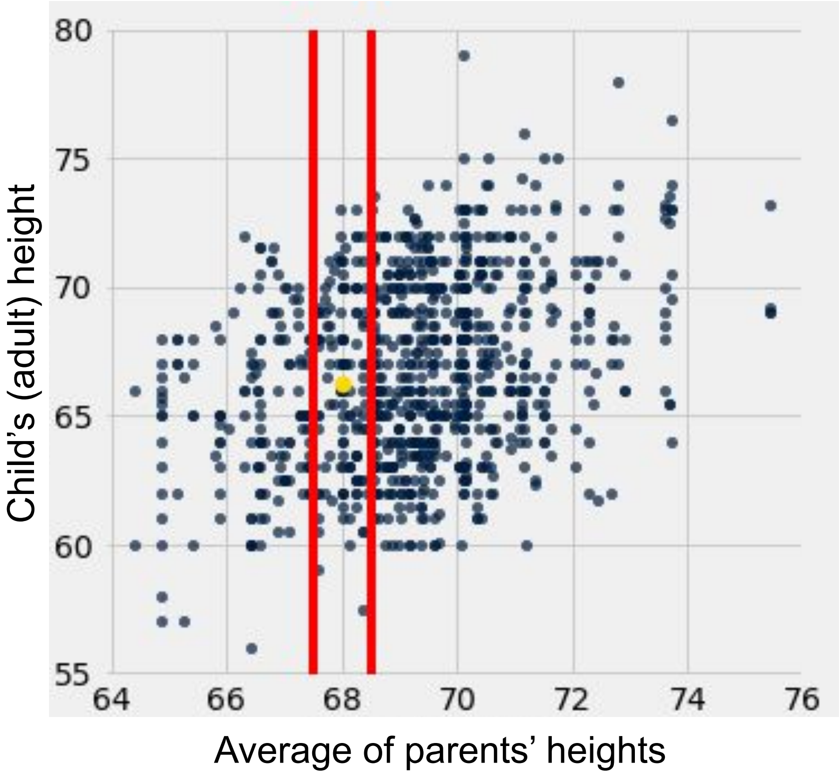
Prediction

Predicting Heights

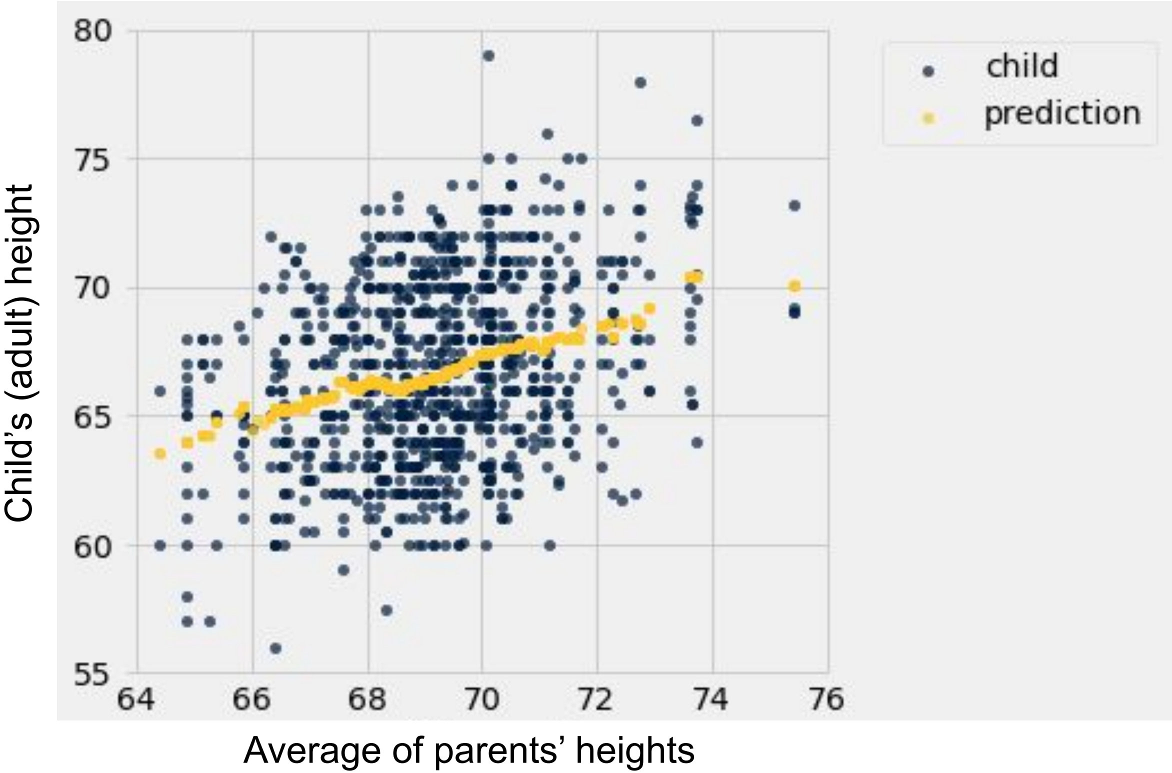


- Oval shaped
- Moderate positive correlation
- How can we predict child height from the parents' average height?

Approach to Prediction



Predicted Heights



Nearest Neighbor Regression

A method for prediction:

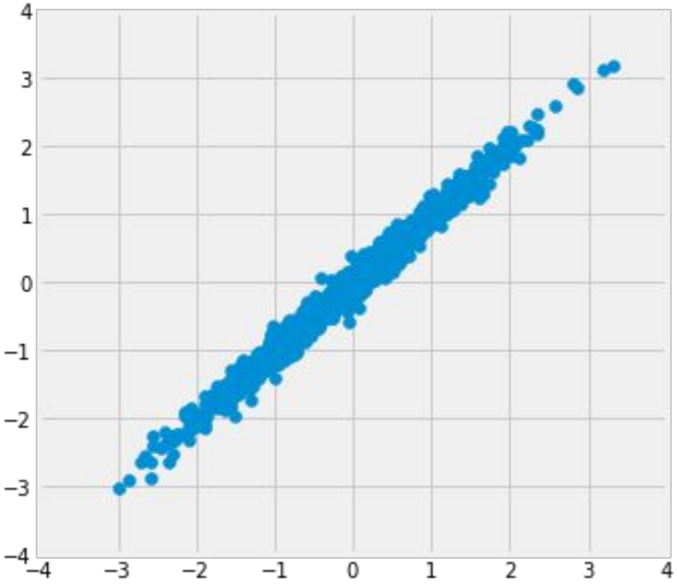
- Group each *x* with similar (nearby) *x* values
- Average the corresponding *y* values for each group

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”.

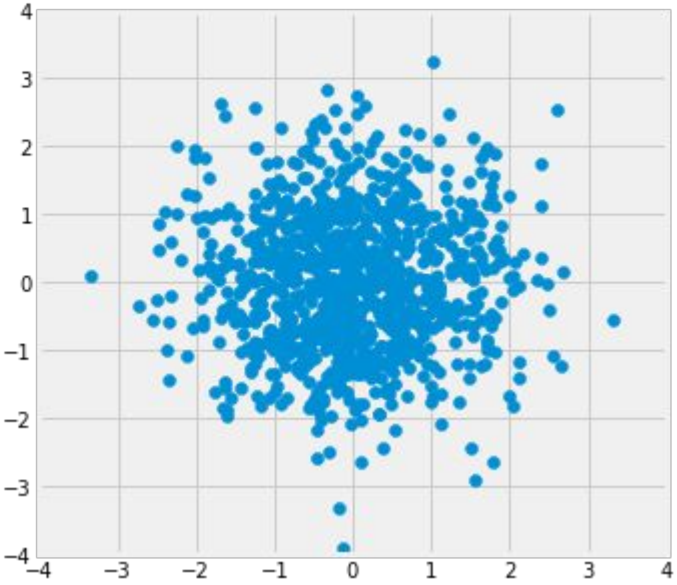
If the association between *x* and *y* is linear, then points in the graph of averages tend to fall on a line.

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 (su)***
- Describing the deviation of *x* from 0 (the average of *x*'s)
- And the deviation of the corresponding *y* from 0 (the average of *y*'s)

On average, *v* deviates from 0 less than *x* deviates from 0

Regression line

$$(\text{estimated } y)_{su} = r \times x_{su}$$

correlation

Not true for all points — a statement about averages