DATA 8

Spring 2020

Lecture 29

Correlation

Regression roadmap

- Today:
  - Association and correlation
- Wednesday
  - 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

Prediction

Guessing the Future

- Based on incomplete information

- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

(Demo)

Association

Two Numerical Variables

- Trend
  - Positive association
  - Negative association
- Pattern
  - Any discernible “shape” in the scatter
  - Linear
  - Non-linear

Visualize, then quantify

(Demo)

Correlation Coefficient

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

Care in Interpretation

Watch Out For ...

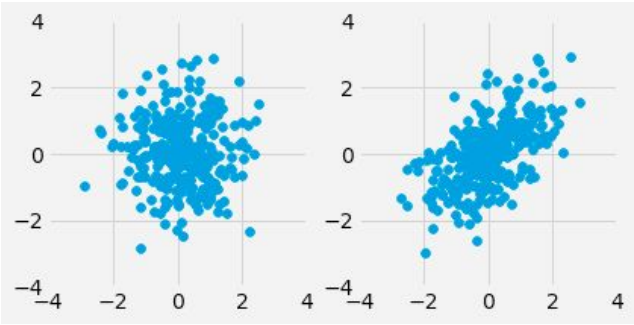
- False conclusions of causation
- Nonlinearity
- Outliers
- Ecological Correlations

(Demo)

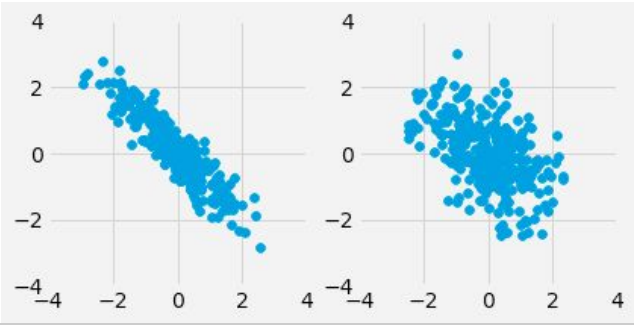
Discussion Question

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

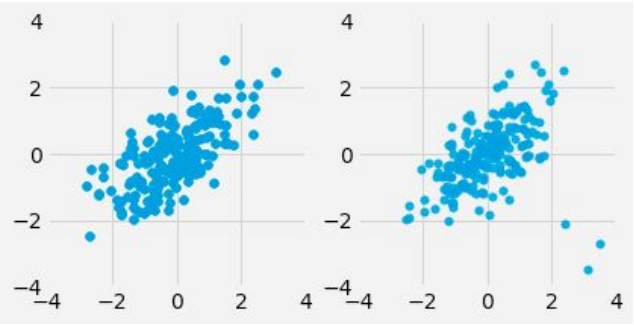
a)



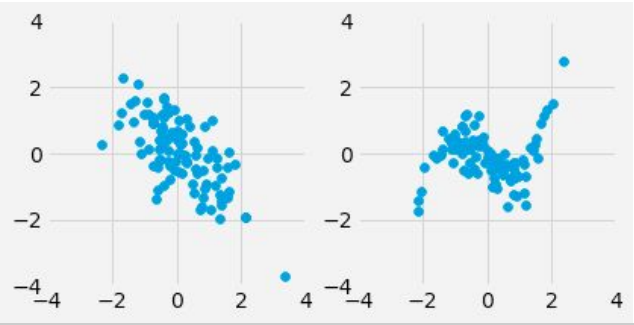
b)



c)



d)



Chocolate and Nobel Prizes

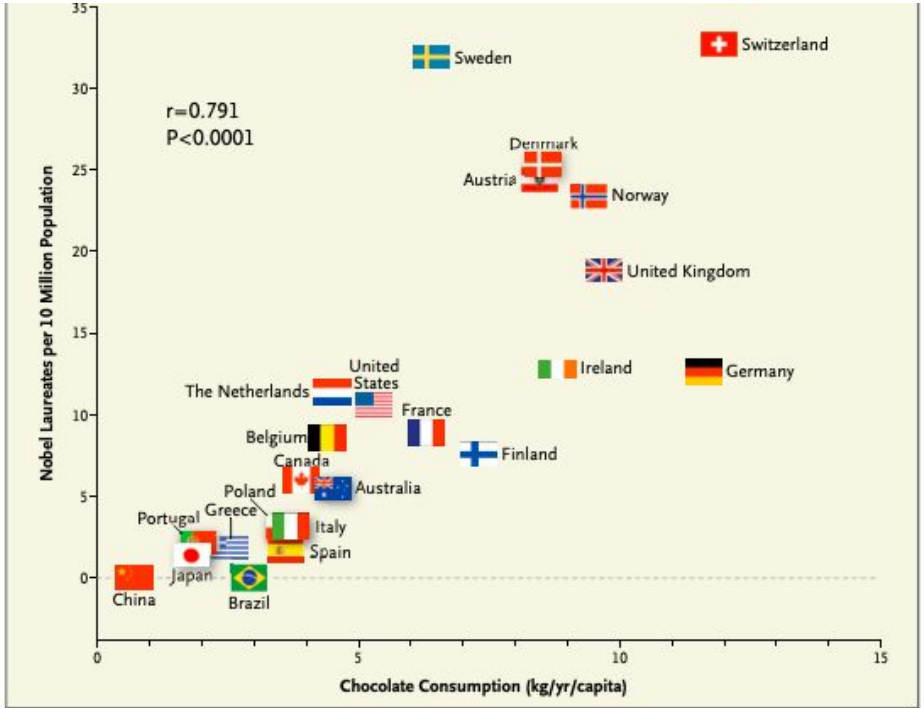


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.