EGM101 – Skills Toolbox

Week 6, Part 2: Correlation (is not Causation)

Covariance

- Recall: variance measures dispersion of a single variable
- Covariance measures the dispersion of two variables
- Pros:
  - Can determine direction of relationship
- Cons:
  - Squared units (like variance)
  - Can't directly compare different variable pairs
  - Hard to determine strength of relationship

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Pearson's Correlation Coefficient

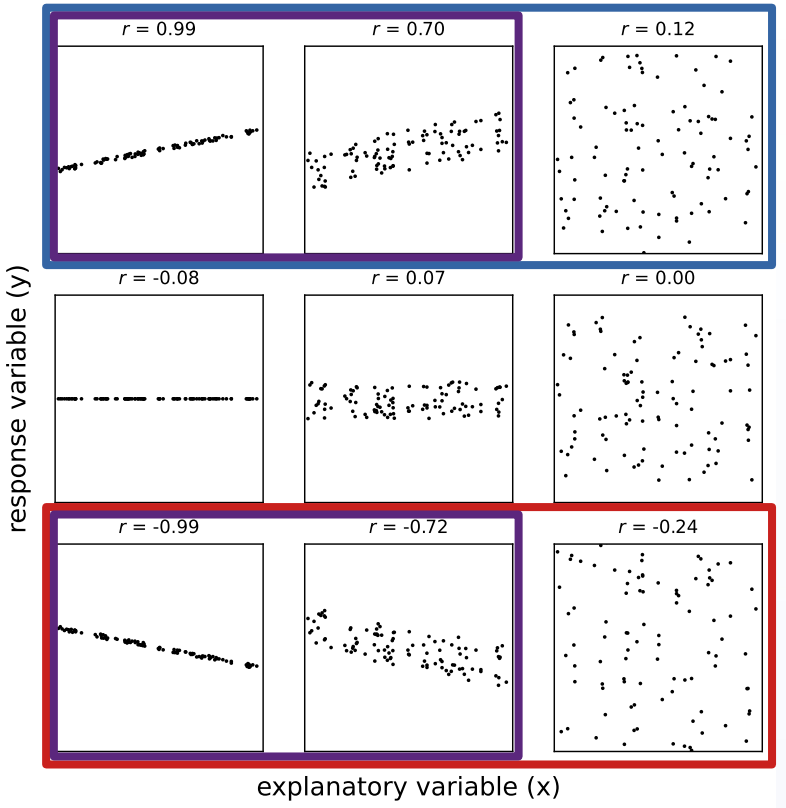
- One solution: divide covariance with standard deviations of x, y
- Pros:
  - No units
  - Values between -1, 1 make comparison easy
  - Tells direction, strength of association/relationship
- Cons:
  - Only tells us about the linear relationship between variables
  - Can be very sensitive to outliers

$$r(x,y) = \frac{\mathrm{cov}(x,y)}{s_x s_y}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

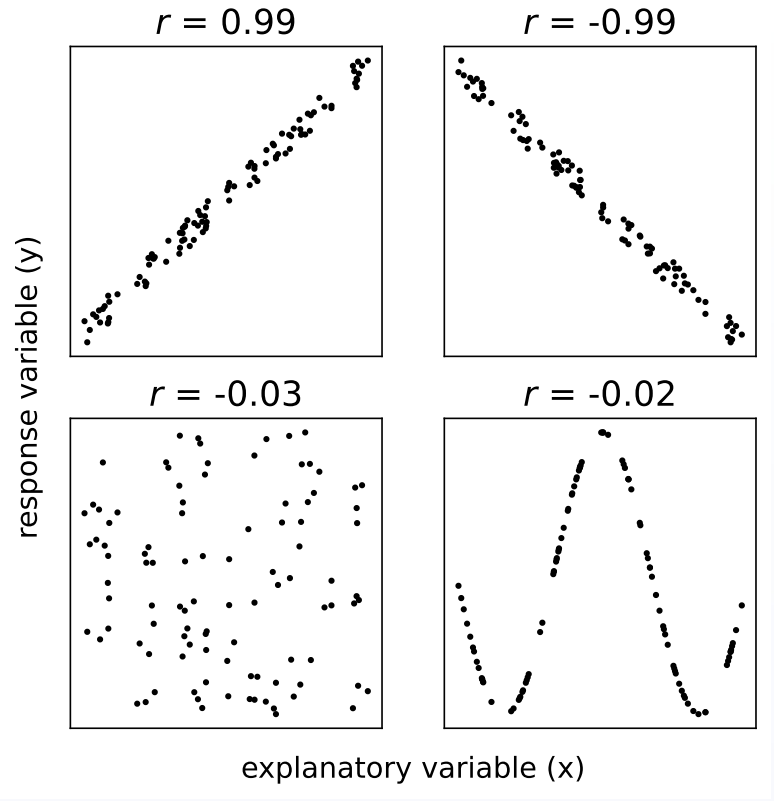
Pearson's Correlation Coefficient

- Direction:
  - $r > 0$: positive association
  - $r < 0$: negative association
- Strength:
  - $|r| \approx 1$: strong linear relationship
  - $|r| = 1$: all points lie on a straight line
  - $|r| \approx 0$: weak linear relationship



Correlation of Nonlinear Relationships

- Remember: Pearson's r only tells us about linear relationship between variables
- In other words, small r does not mean there is no association!



Spearman's Rank Correlation Coefficient

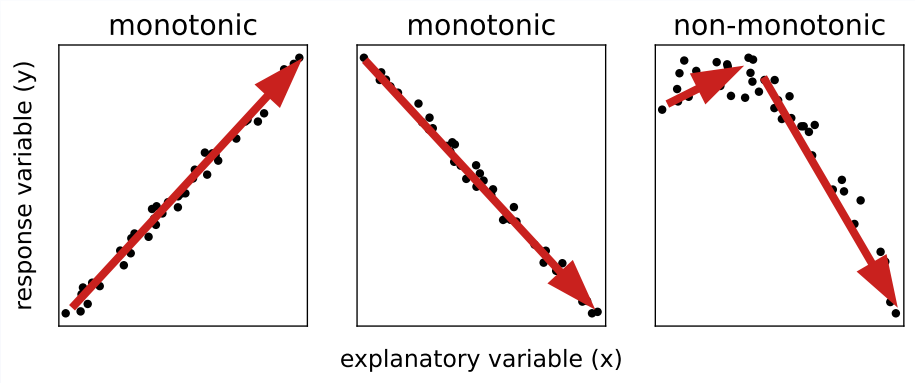
- Uses the difference in rank, R, of pairs of data
- Assesses whether variables are monotonically related
- Pros:
  - Don't need actual values (just ranks)
  - Data can be numeric (continuous, discrete) or ordinal (i.e., non-numeric)
  - Less sensitive to outliers
- Cons:
  - Ties can be tricky (can't use simplified formula)

x	8	5	2	7
$R(x)$	1	3	4	2

y	3	5	4	9
$R(y)$	4	2	3	1

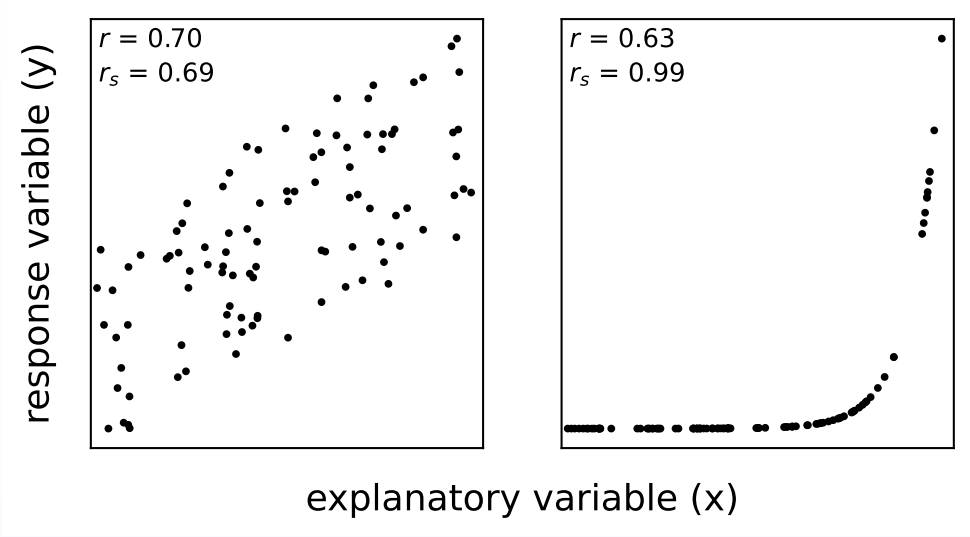
$$r_s = \frac{\operatorname{cov}\left(R(x), R(y)\right)}{s_{R(x)} s_{R(y)}}$$

$$\approx 1 - \frac{6\sum_{i=1}^{n}\left(R(x_i) - R(y_i)\right)^2}{n\left(n^2-1\right)}$$



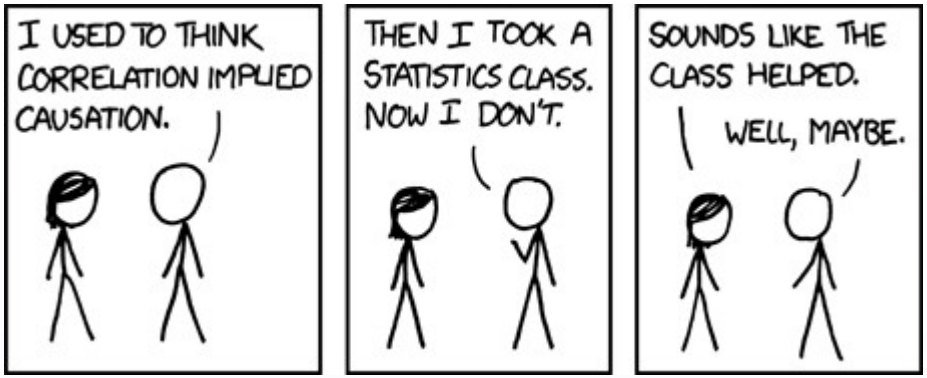
Pearson vs Spearman Correlation

- r_s is usually a good approximation of r
- For some nonlinear relationships it works much better:
  - Exponential
  - Logarithmic

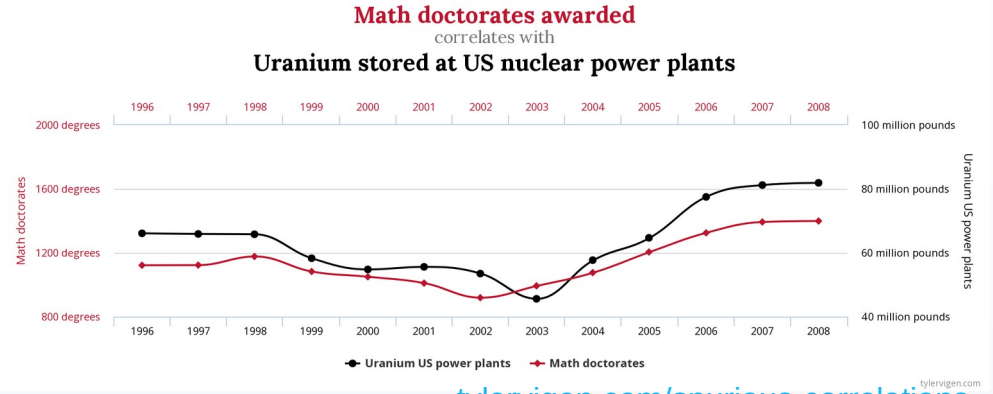


Correlation is not Causation!

- Remember: correlation tells us about the association between two variables
- Cannot determine which is explanatory/independent, response/dependent
  - In other words, correlation <u>cannot</u> tell us anything about causality



xkcd.com/552



tylervigen.com/spurious-correlations

Summary

- Correlation:
  - Helps describe the (linear) relationship between variables.
  - Can be determined for numeric or ordinal data
  - Is not causation.
- Correlation doesn’t tell us about causality.
- Even if it looks like it might be, correlation is still not causation.

Additional Resources

- Illowsky and Dean, Chapter 12.3
- Caswell, Chapter 9.5 – 9.7
- Weiss, Chapter 14.4
- Huff, “Post Hoc Rides Again” (Chapter 8)
- Bergstrom and West, “Causality” (Chapter 4)
- tylervigen.com/spurious-correlations