Statistics with two variables

MOL518, Lecture #12

Overview and notation

- For this lecture, we will consider two variables, X and Y
- You can think of X and Y as molecular concentrations:
  - The expression levels of two genes (RNA or protein)
  - The RNA and protein levels of the same gene
  - Metabolite, or lipid levels
- We will examine both linear and nonlinear relationships between X and Y
- Aside: notation in statistics is very confusing, certain letters get overused

Why does this matter?

- Can help us determine if X regulates Y (activation or repression)
- Alternatively, X and Y could be co-regulated by some other factor Z
- Each of these scenarios generates a testable hypothesis!
- Correlation and regression also help with testing a new method: how does it compare to the state of the art?

Outline

- Pearson correlation
- Linear regression and least-squares fitting
- Logistic regression
- Rank-based correlation

Outline

- Pearson correlation
- Linear regression and least-squares fitting
- Logistic regression
- Rank-based correlation

Defining covariance mathematically using the expected value

- Define the expected value E

- $E[X] = \sum_{i=1}^{n} x_i p_i$, where n is the sample size
- x_i are possible outcomes of X and p_i are there corresponding probabilities

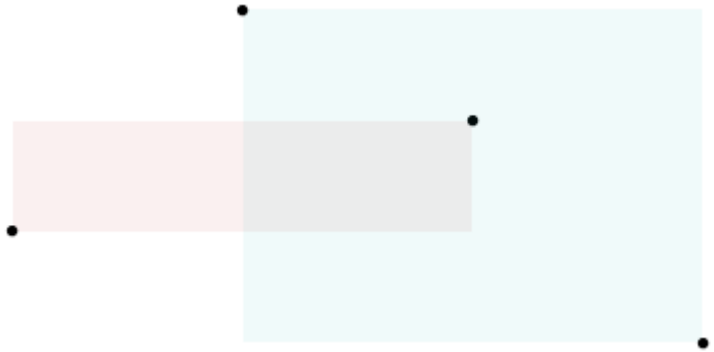
- Then $\mathrm{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

- A special case: $\mathrm{cov}(X, X) = \mathrm{var}(X) = \sigma_X^2$

A visual explanation of covariance

- Make a scatterplot of your data

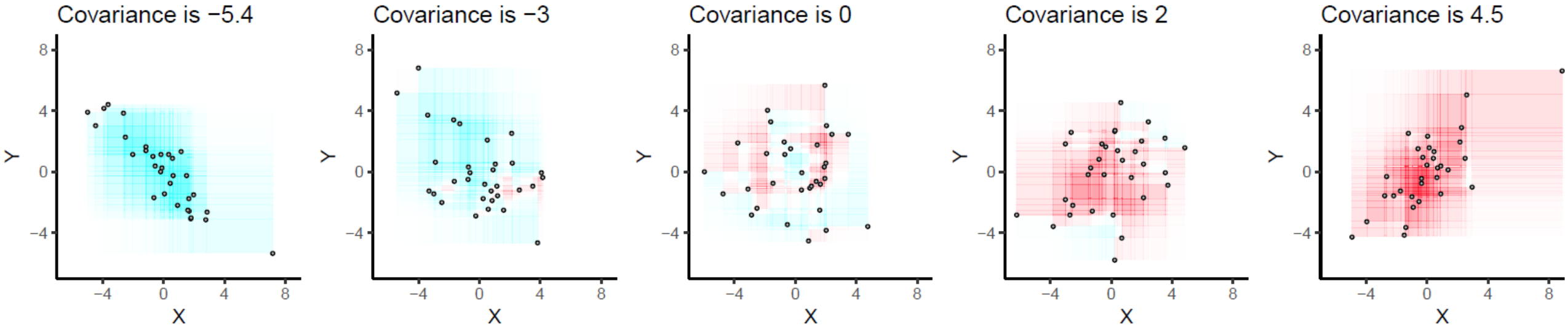
- Draw all possible rectangles through the data
- Color the rectangles:
- Red = upward sloping
- Blue = downward sloping



- Treat red as positive values and blue as negative
- The covariance is the net amount of red in the plot
- ***Technically this measures 2x the covariance

Source: https://stats.stackexchange.com/questions/18058/how-would-you-explain-covariance-to-someone-who-understands-only-the-mean

A visual explanation of covariance



- Covariance is proportional to the scale of the plot (because it's proportional to area)
- Covariance is great for linear relationships
- Covariance is sensitive to outliers (because they create many large rectangles)

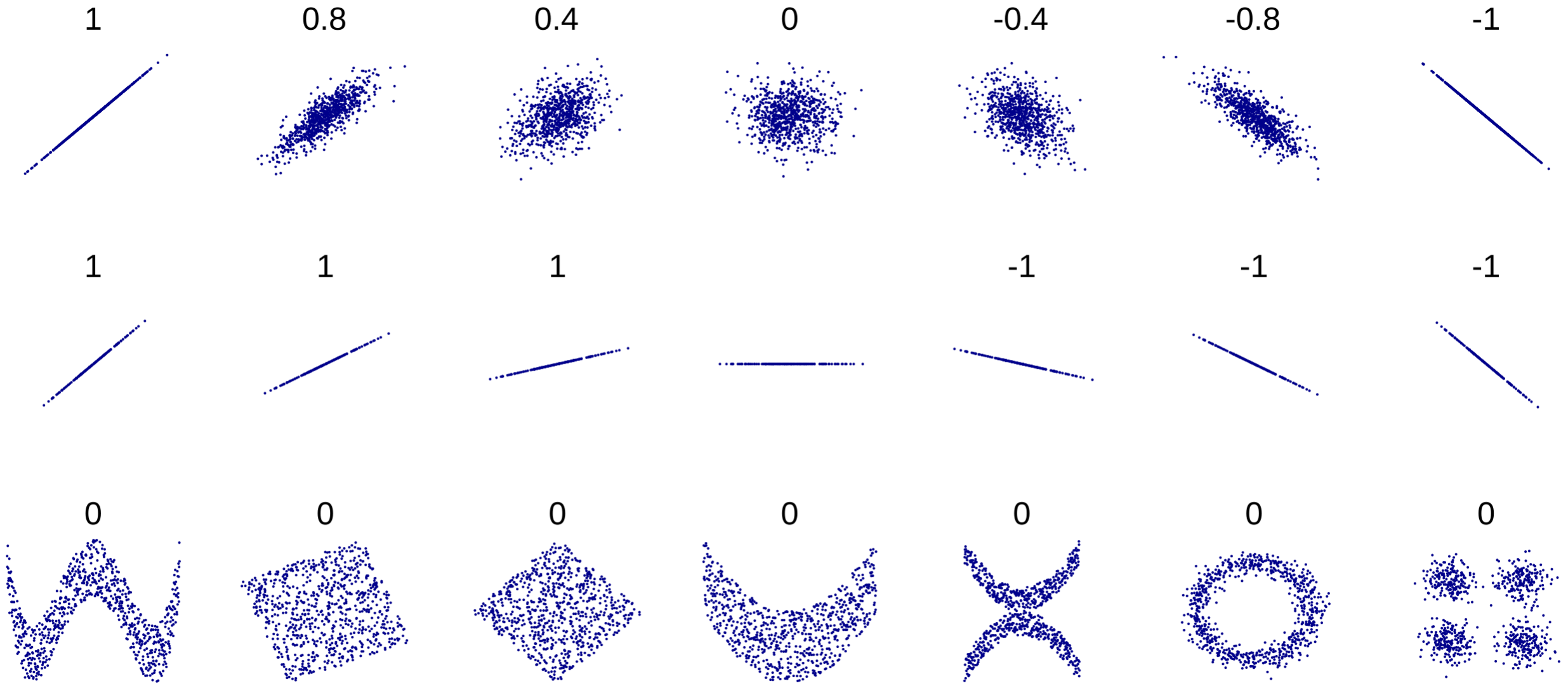
Source: https://stats.stackexchange.com/questions/18058/how-would-you-explain-covariance-to-someone-who-understands-only-the-mean

Pearson correlation is a normalized covariance

- Are two variables (X and Y) linearly related to one another?
- In the real world, we can estimate the correlation r using by estimating the covariances and variances of X and Y
- $r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\text{estimated cov}(X,Y)}{\text{estimated } \sigma_X \sigma_Y}$; here, n is the sample size
- Due to normalization, r can vary from -1 to $+1$
- This is important so that r does not depend on the scale of the data!!

Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Pearson correlation coefficient examples



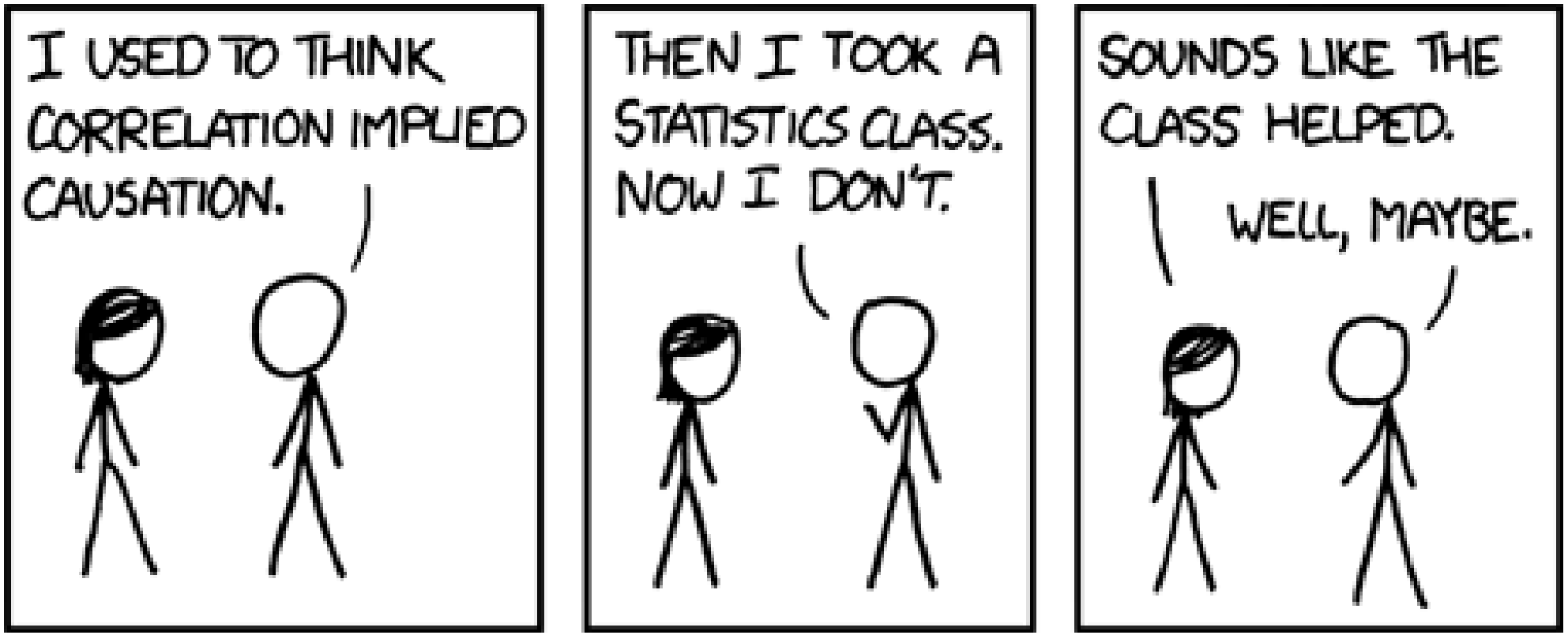
Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#/media/File:Correlation_examples2.svg

Some correlation pitfalls

1. $r = 1$ is very unlikely to happen by chance in biology,
   - Example: temperature in Fahrenheit vs. Celsius
2. Many relationships between two biological variables are not linear!
3. Repeated measurements or time series are **not independent**
4. When in doubt, plot the data – scatterplots are your friend ☺

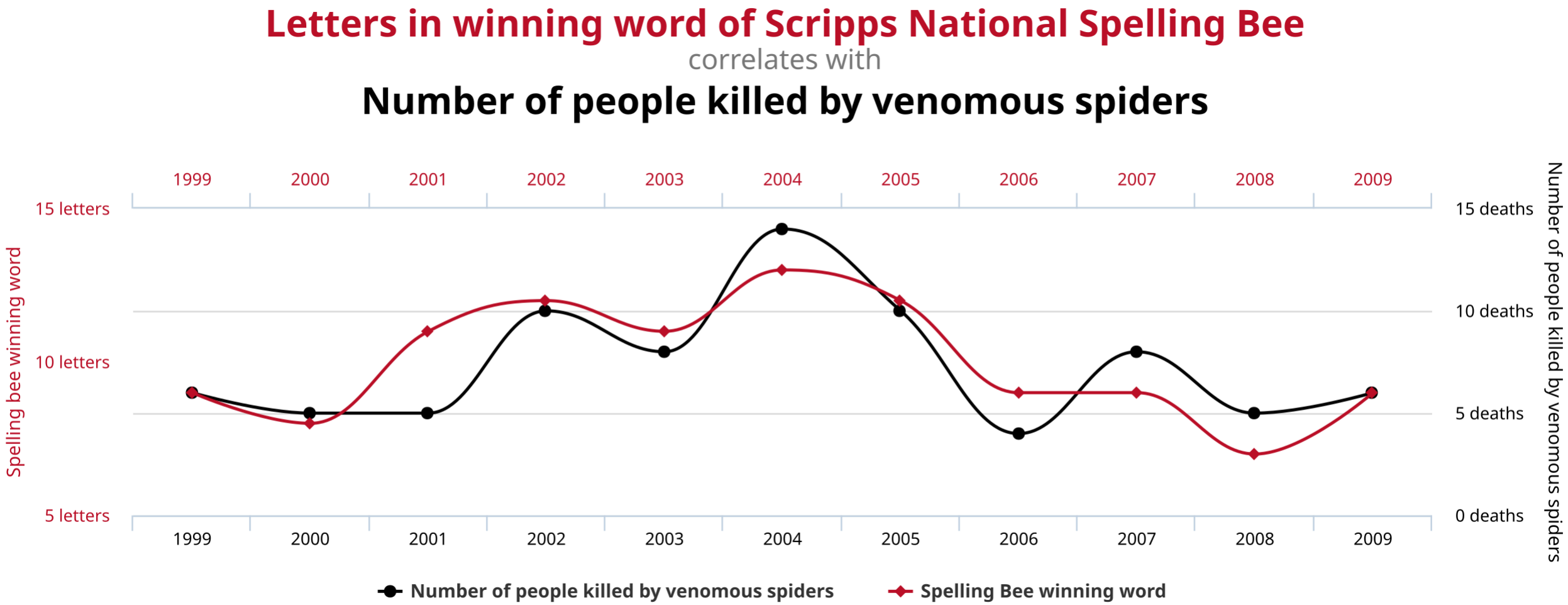
Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#/media/File:Correlation_examples2.svg

Correlation does not imply causation



Source: https://xkcd.com/552/

Another example of spurious correlation



Source: https://en.wikipedia.org/wiki/Multiple_comparisons_problem

Outline

- Pearson correlation
- Linear regression and least-squares fitting
- Logistic regression
- Rank-based correlation

Why do we want to do linear regression?

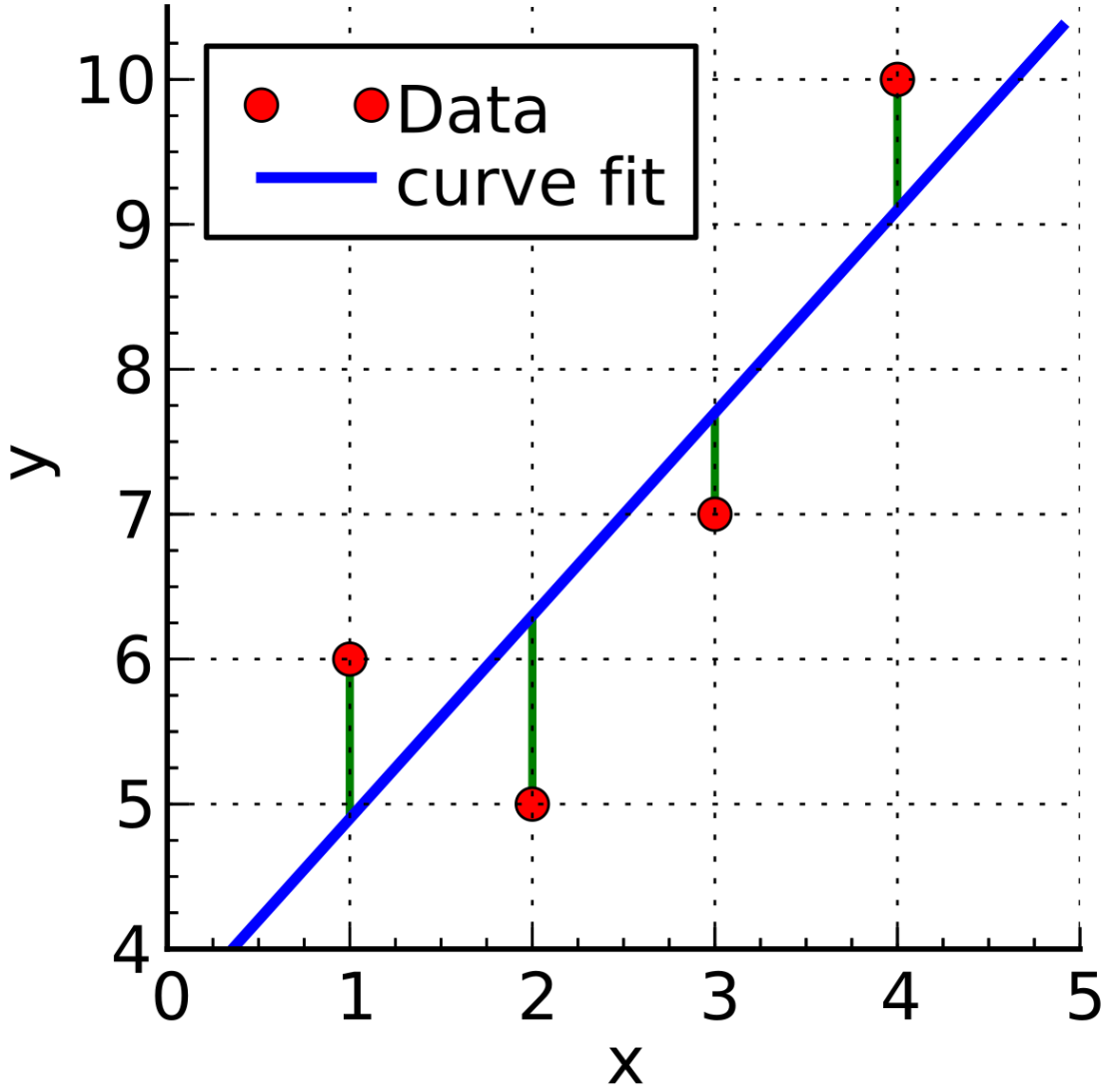
- In general, so that we can predict some variable Y based on a known variable X
- This dates all the way back to Isaac Newton in 1700, who tried to predict the timing of equinoxes
- The least squares method was first performed by Legendre and Gauss in the early 1800s

How does linear regression work?

- Draw a regression line through the data (x_i, y_i)
- We are trying to find the line that best fits the data: $\hat{Y} = a + bX$
- This sounds easy but is tricky in practice

- Typically, this is done by minimizing the sum of the squared vertical distance between the line and the data:
- $\sum_{i=1}^{n} \left(y_i - \hat{Y}\right)^2$

Assumptions associated with linear regression

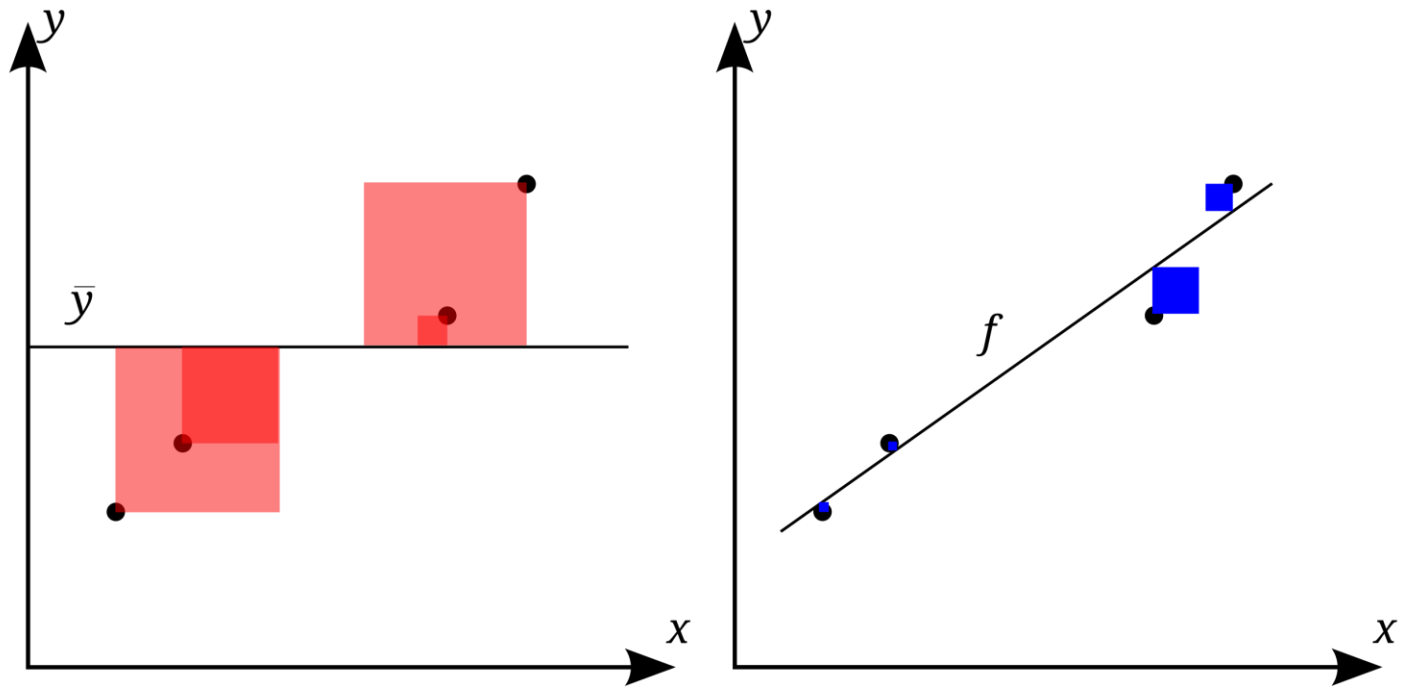


- X values are fixed (no error)
- Linearity
- Constant variance
- Errors in Y are uncorrelated with one another
- No outliers

Source: https://en.wikipedia.org/wiki/Linear_regression

How do we know if the fit is good

- Calculate a coefficient of determination R^2, which is the proportion of variation we can predict
- It turns out that R^2 is the square of the Pearson correlation coefficient!
- As such, it scales between 0 and 1

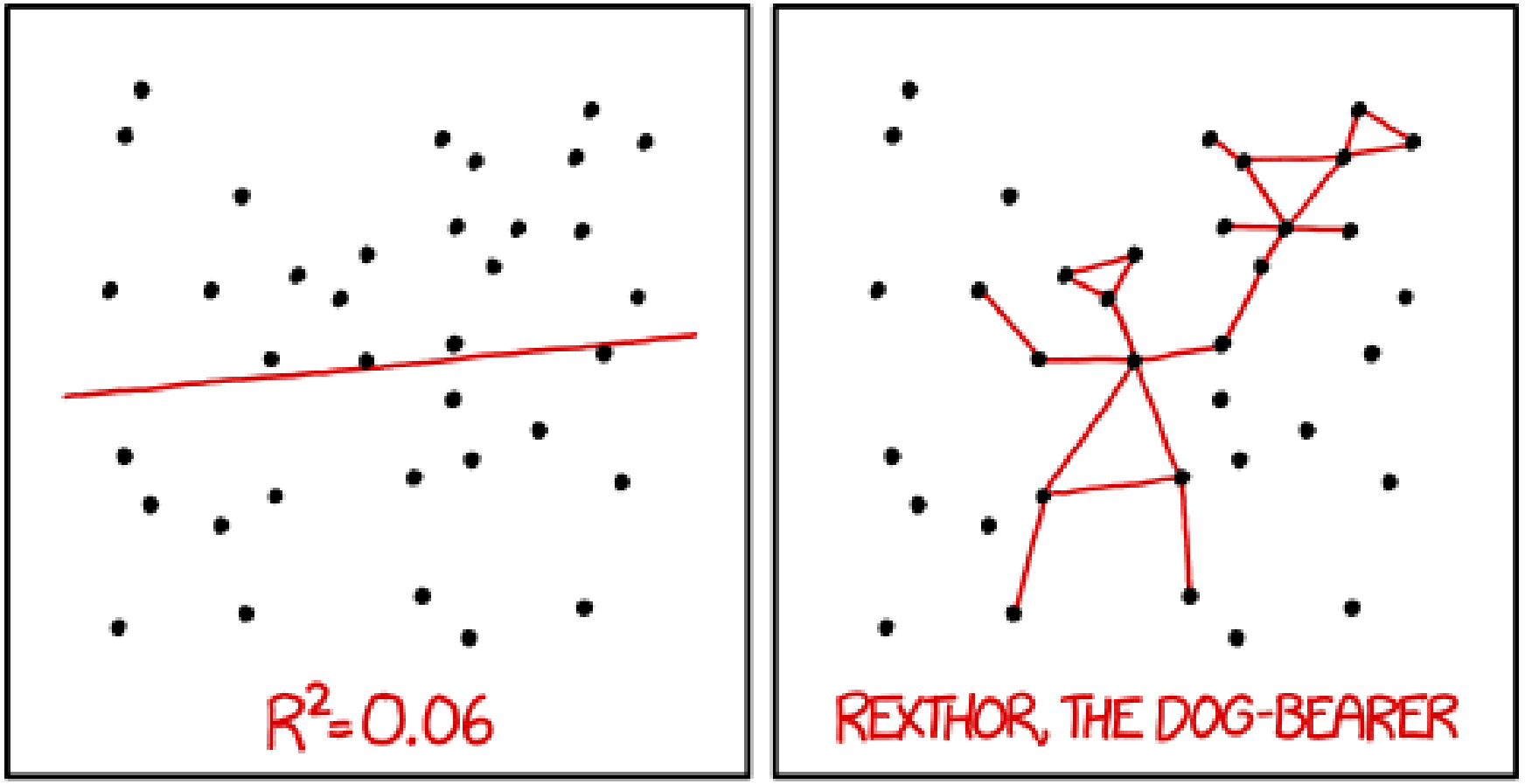


Source: https://en.wikipedia.org/wiki/Coefficient_of_determination

Comparing Pearson correlation and linear regression

- R^2 from a simple linear regression is the square of Pearson's r!!!
- Linear regression allows you to predict Y from X
- Pearson correlation between X and Y is symmetric
- Linear regression between X and Y is **not symmetric** because we are trying to predict Y from X or vice-versa

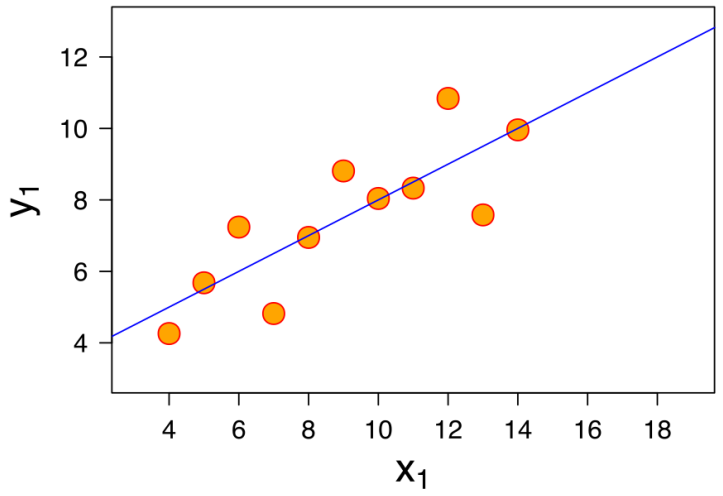
Some limitations of linear regressions

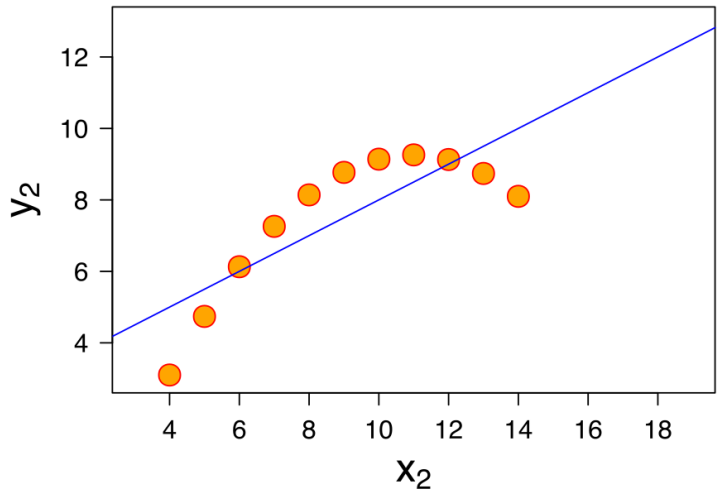


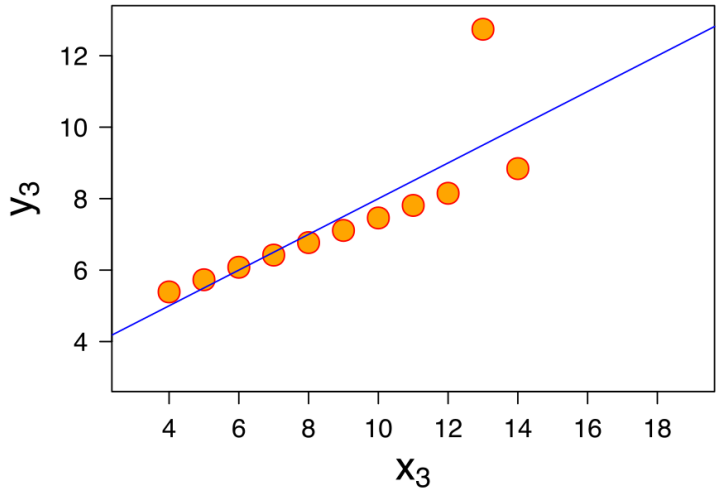
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

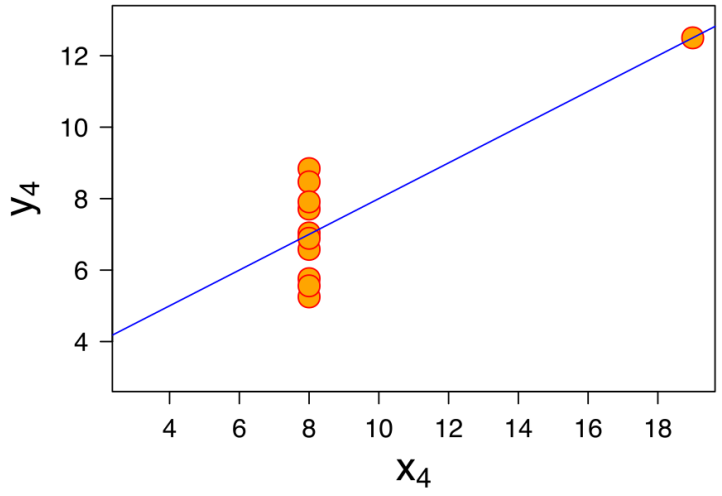
Source: https://xkcd.com/1725/

Limitations of linear analyses: Anscombe's quartet









Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x: s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y: s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression: R^2	0.67	to 2 decimal places

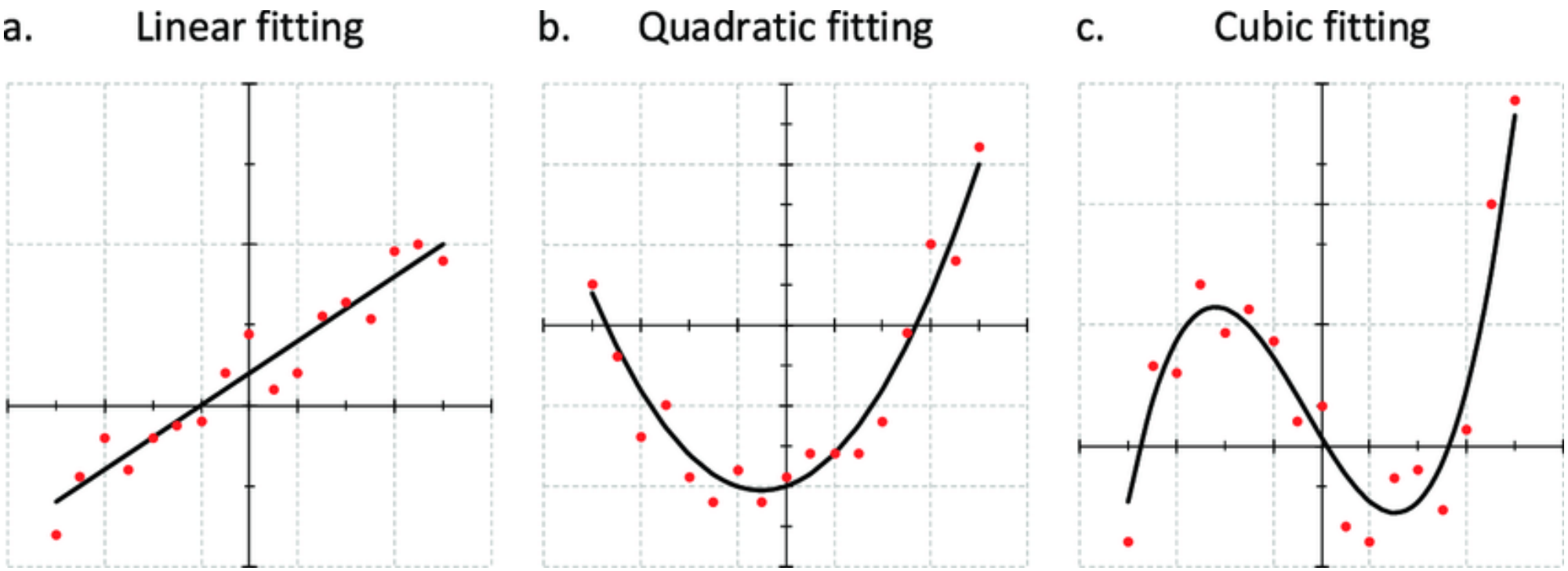
Source: https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Help! My data are nonlinear! What do I do?

- Fit a more complicated function:
  - Polynomial
  - Exponential
  - Sine wave

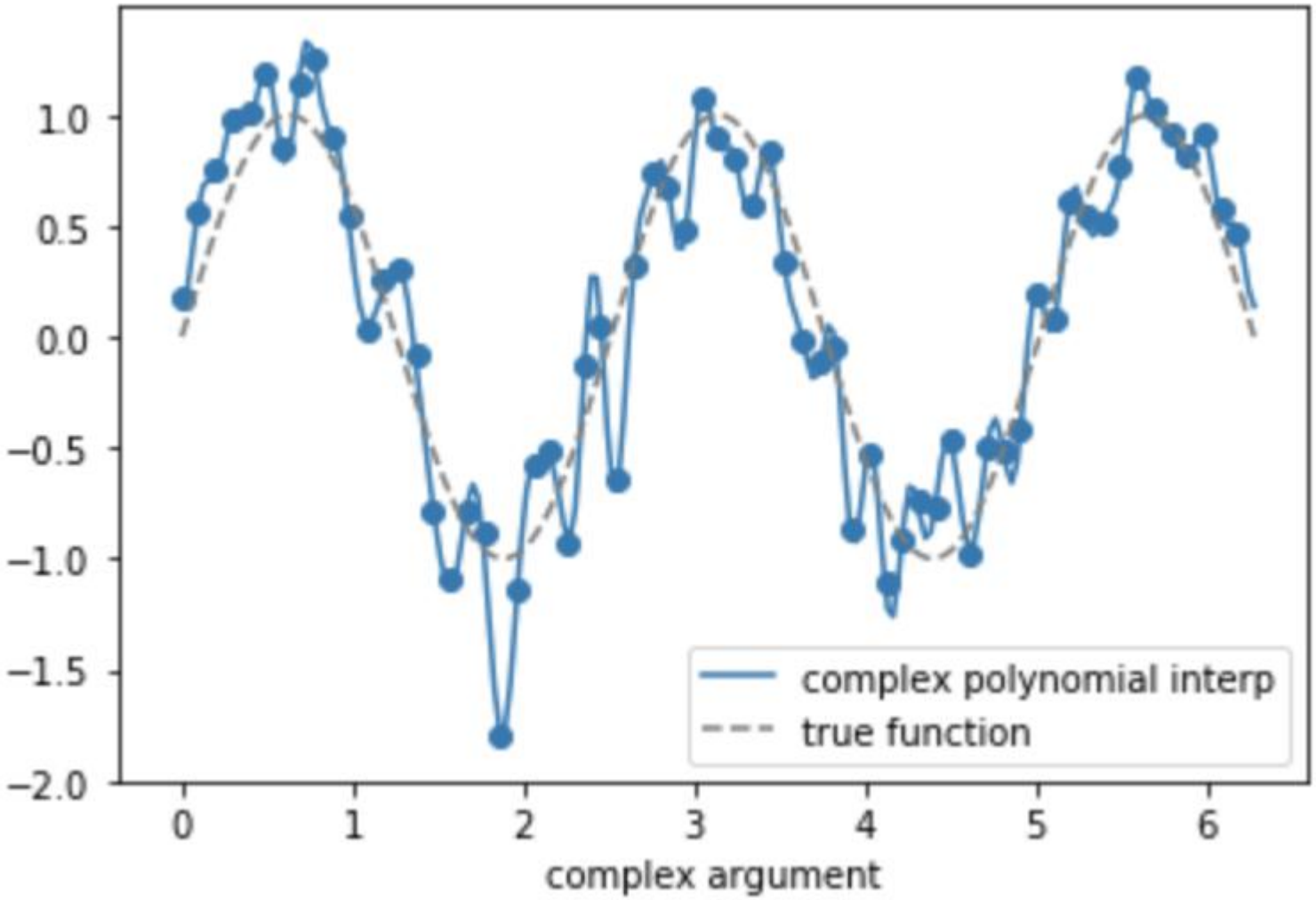
- Rank-based correlation

Polynomial regression examples



Source: https://www.sciencedirect.com/science/article/pii/S0141029618313257?via%3Dihub

Overfitting is a challenge when doing computational work!



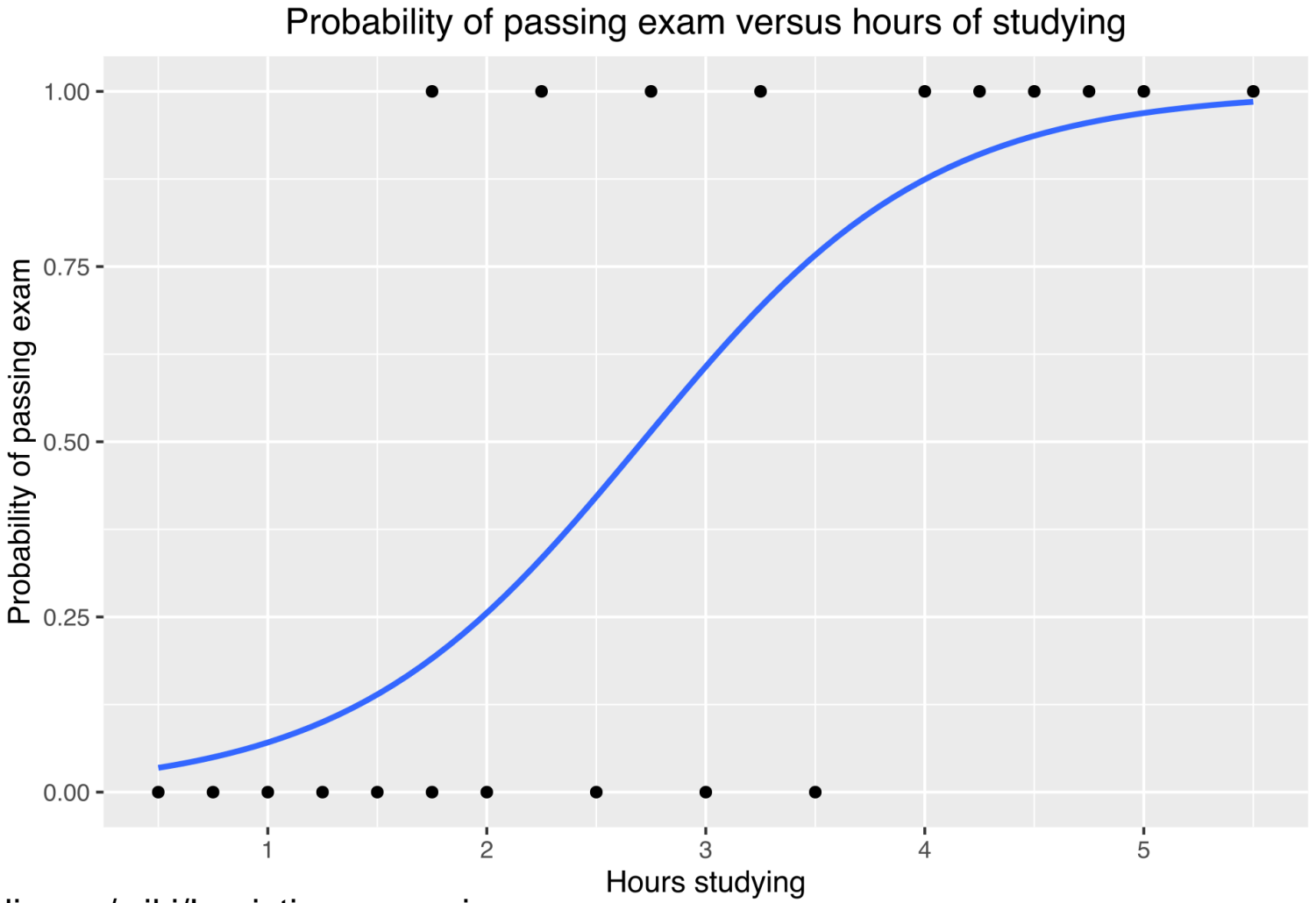
Source: https://stats.stackexchange.com/questions/560383/how-can-we-explain-the-bad-reputation-of-higher-order-polynomials

Outline

- Pearson correlation
- Linear regression and least-squares fitting
- Logistic regression
- Rank-based correlation

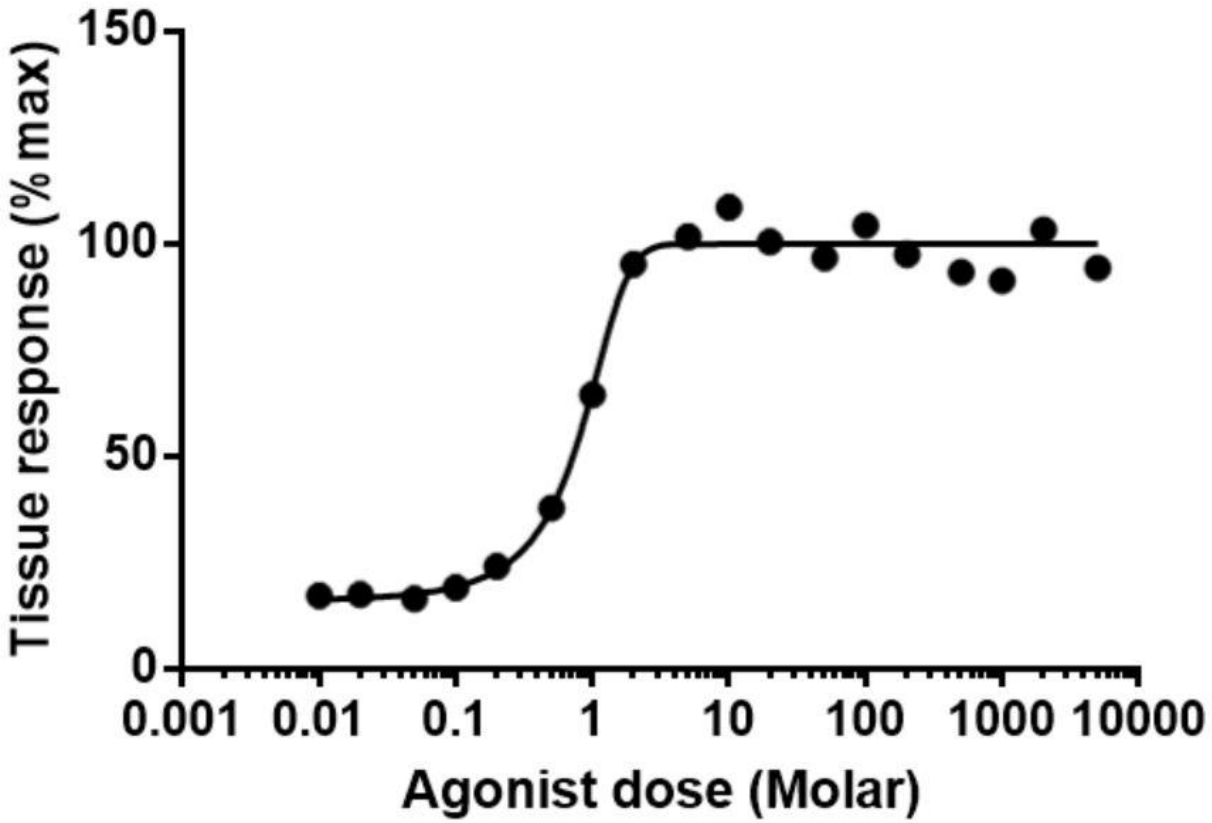
What is logistic regression?

- Logistic regression is a special case when the Y axis is a probability (or binary)



Source: https://en.wikipedia.org/wiki/Logistic_regression

Logistic regression is useful for dose-response curves



We will discuss this in more detail later in the course!

Source: https://en.wikipedia.org/wiki/Dose%E2%80%93response_relationship

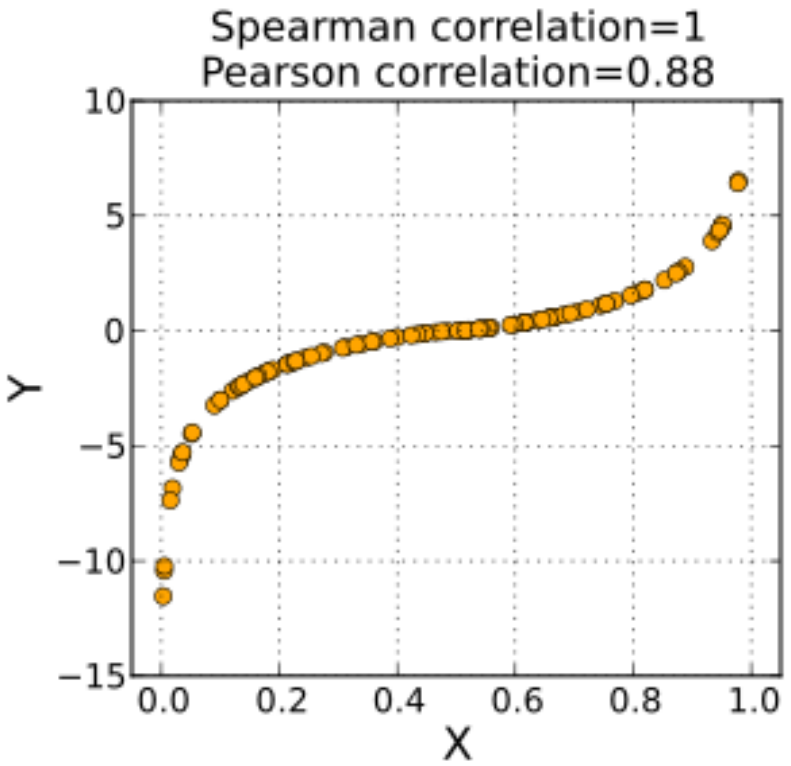
Outline

- Pearson correlation
- Linear regression and least-squares fitting
- Logistic regression
- Rank-based correlation

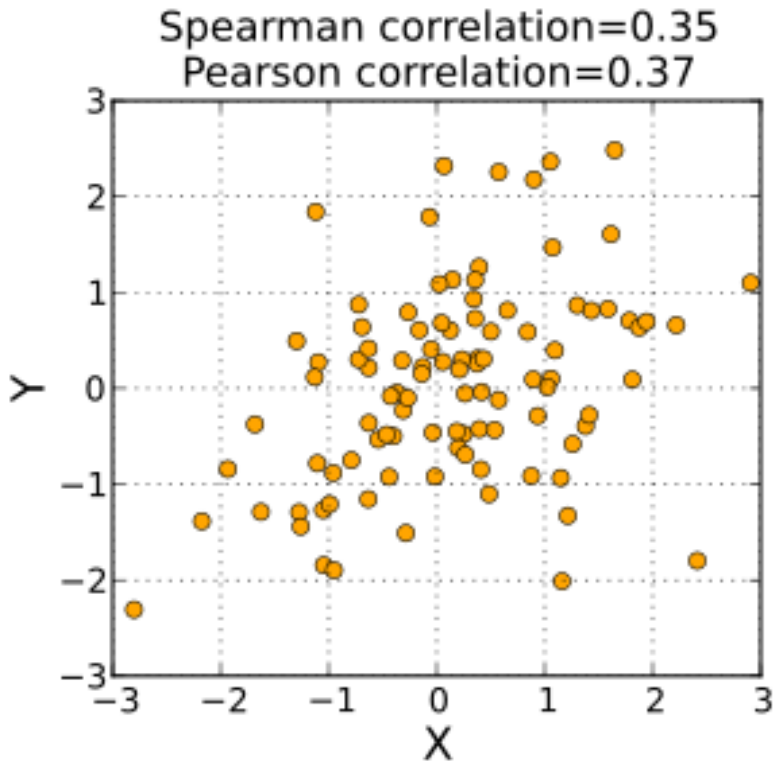
Spearman's ρ (rank-based correlation)

- Like Pearson correlation, but for ranks rather than actual values:
- Let $R[X]$ be the ranks of X, and $R[Y]$ be the ranks of Y
- $X = 1, 4, 7; Y = 4,6, 10$
- $R[X] = 1,2,3; R[Y] = 1,2,3$
- $\rho = \frac{cov(R[X],R[Y])}{\sigma_{R_X}\sigma_{R_Y}}$
- This is the same calculation we did for Pearson correlation, but with ranks

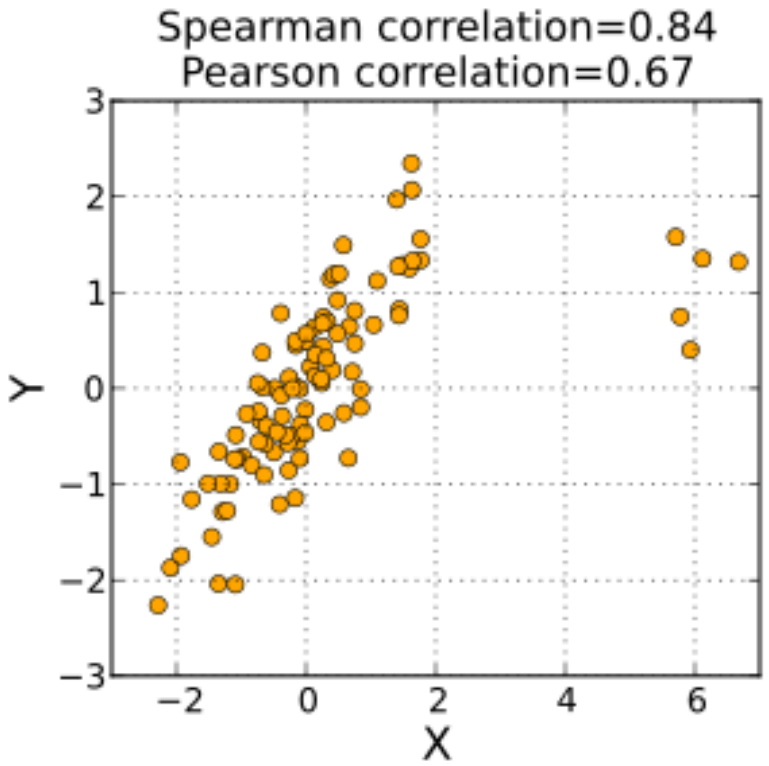
Spearman vs. Pearson correlation



Spearman correlation is **nonlinear**



Sometimes both methods are similar



Spearman correlation is **less sensitive to outliers**

Source: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

When should you use Spearman correlation?

1. When there are outliers in your data

2. When your data are nonlinear

Extra Slides