YData: An Introduction to Data Science

Lecture 31: Least Squares

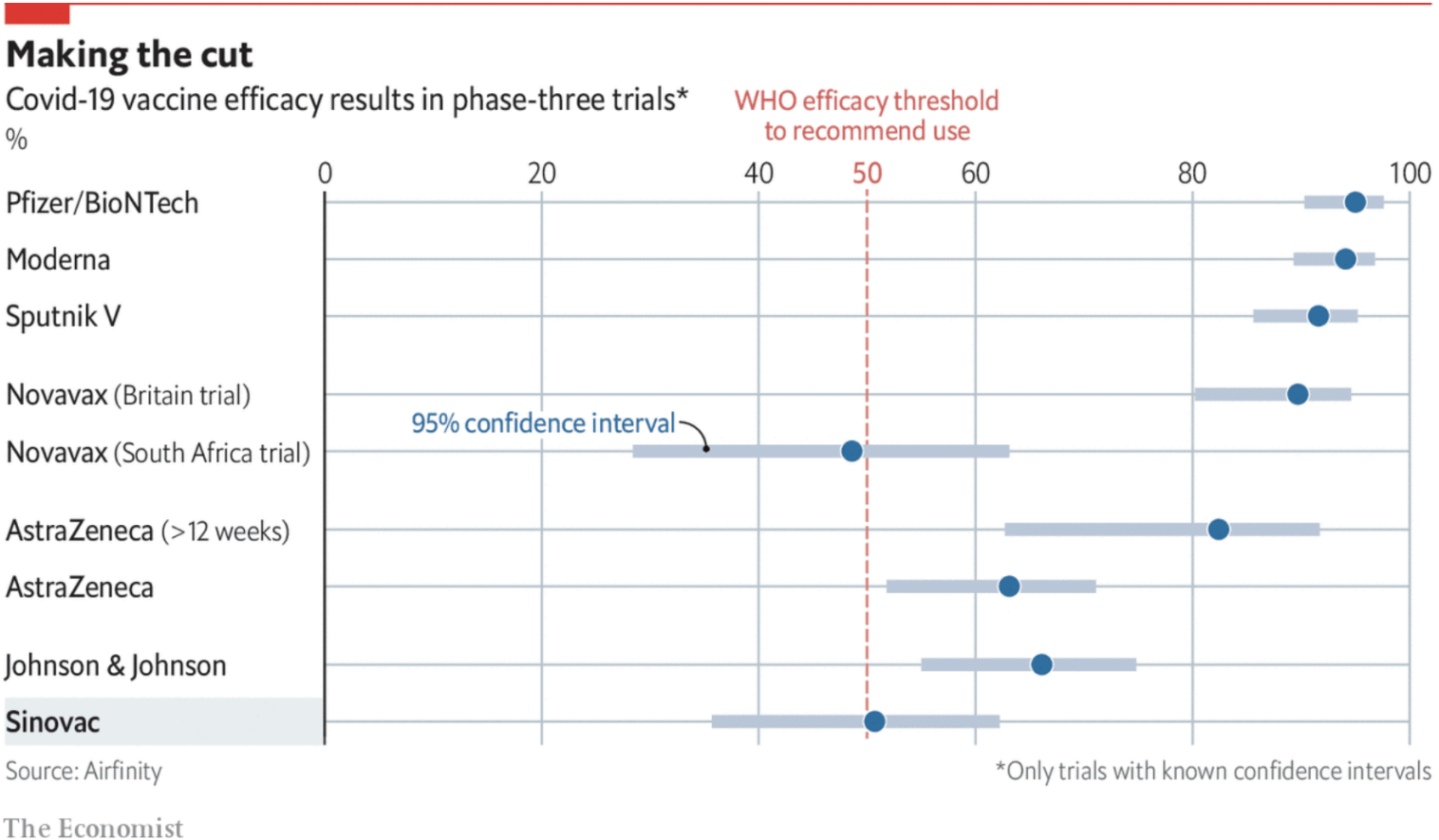
Elena Khusainova & John Lafferty
Statistics & Data Science, Yale University
Spring 2021

Credit: data8.org

Announcements

- Project 2 due today—both partners submit
- Assignment 10 posted; due Thursday 4/22
- Project 3 posted; checkpoint 4/23; due 4/30
- News: Lowest Project score will be dropped. This effectively means Project 3 is optional
- Please see announcment about final exam date and time: Open on Gradescope at 2pm EDT, May 18; Closed at 9pm EDT, May 19.

The Economist: Daily Graphic



Outline for Today

- Review regression in terms of correlation
- Discuss regression in terms of least squares
- Optimization approach to regression
- A peak at nonlinear regression

Linear Regression (Review)

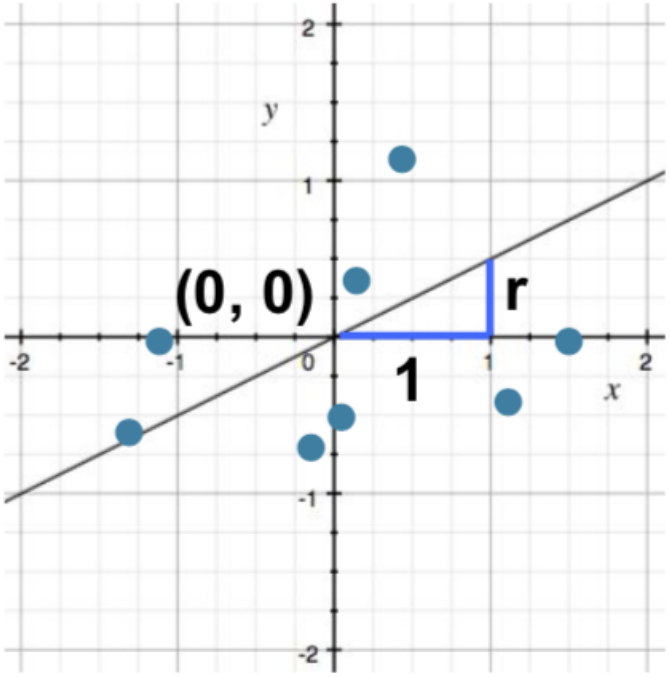
Regression Estimate

To find the regression estimate of y:

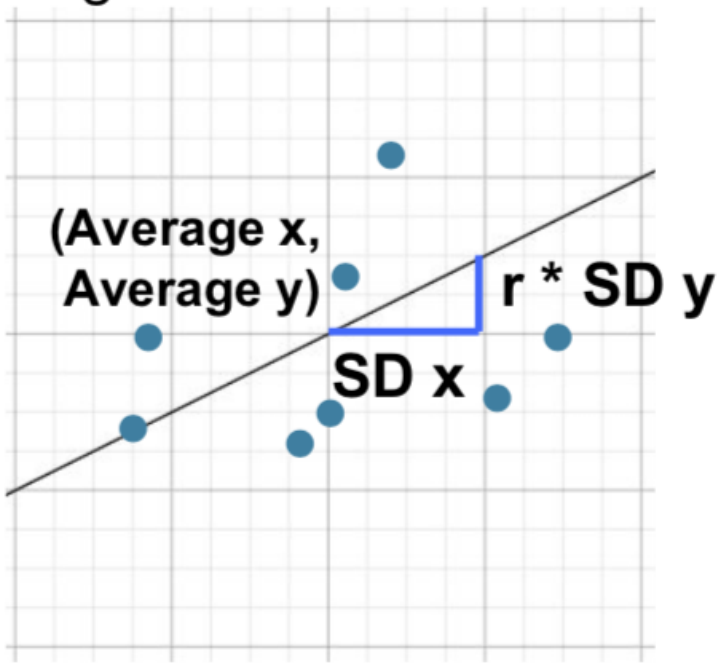
- Convert the given x to standard units
- Multiply by r
- That's the regression estimate of y, but:
  - It's in standard units
  - So covert it back to the original units of y

Regression Line

Standard Units



Original Units



Slope and Intercept

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

$$\textbf{slope of regression line} = r \cdot \frac{\text{SD of y}}{\text{SD of x}}$$

$$\textbf{intercept of regression line} = \text{average of } y \text{ - slope}\cdot\text{average of } x$$

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 a typical 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)

Least Squares

Error in Estimation

- **error = actual value − estimate**
- Typically, some errors are positive and some negative
- To measure the rough size of the errors
  - **square** the **errors** to eliminate cancellation
  - take the **mean** of the squared errors
  - take the square **root** to fix the units
  - **root mean square error** (rmse)

(DEMO)

Least Squares Line

- Minimizes the root mean squared error (rmse) among all lines
- Equivalently, minimizes the mean squared error (mse) among all lines
- Names:
  - "Best fit" line
  - Least squares line
  - Regression line

Numerical Optimization

- Numerical minimization is approximate but effective
- Much of machine learning is based on numerical minimization
- If the function `mse(a, b)` returns the mse of estimation using the line "estimate = ax + b",
  - then `minimize(mse)` returns array `[`a_0`,` b_0`]`
  - a_0 is the slope and b_0 the intercept of the line that minimizes the mse among lines with arbitrary slope a and arbitrary intercept b (that is, among all lines)

(DEMO)