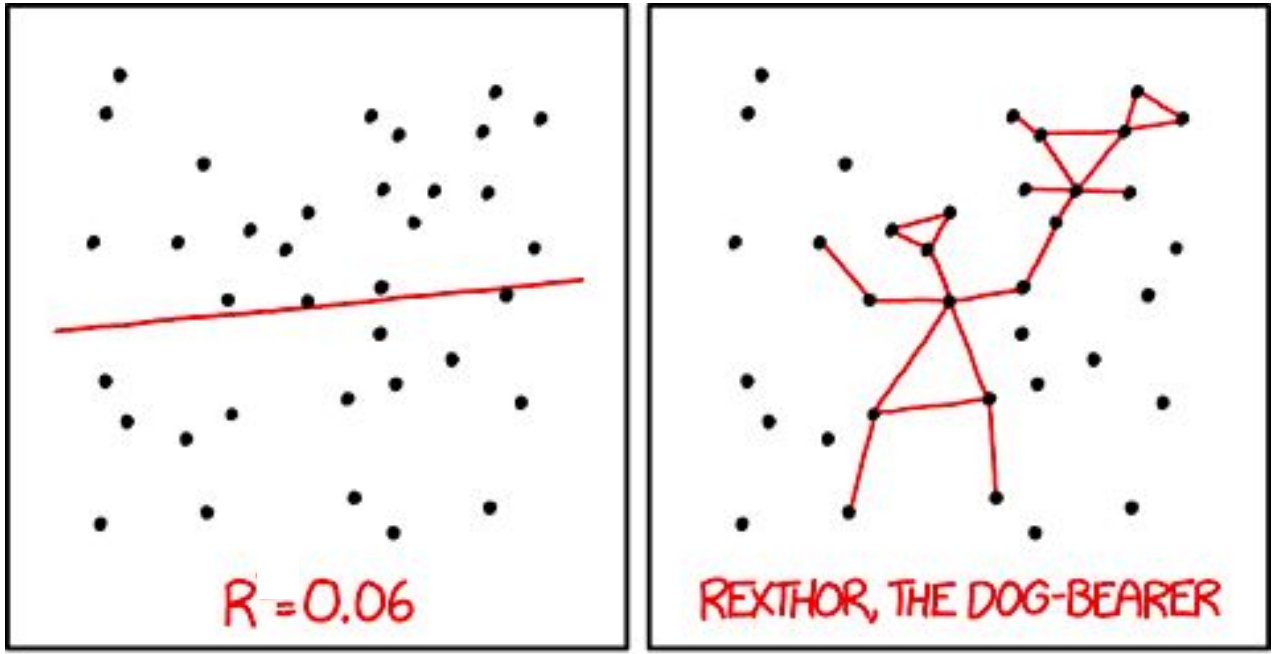
DATA 8

Spring 2020

Lecture 31

Least Squares



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: xkcd.com/1725

Regression roadmap

- Monday:
  - Association and correlation
- Wednesday:
  - Prediction, scatterplots and lines
- **Today:**
  - **Least squares: finding the “best” line for a dataset**
- Next Monday:
  - Residuals: analyzing mistakes and errors
- Next Wednesday:
  - Regression inference: understanding uncertainty

Linear Regression

Prediction

Goal: Predict *y* using *x*

Examples:

- Predict # *hospital beds available* using *air pollution*
- Predict *house prices* using *house size*
- Predict # *app users* using # *app downloads*

Regression Estimate

Goal: Predict *y* using *x*

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 convert it back to the original units of *y*

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 an 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)

Linear Regression

A statement about x and y pairs

- Measured in *standard units*
- Describing the deviation of x from 0 (the average of x's)
- And the deviation of y from 0 (the average of y's)

On average, y deviates from 0 less than x deviates from 0

Regression Line

$$y_{(\text{su})} = r \times x_{(\text{su})}$$

Correlation

Not true for all points — a statement about averages

Slope & Intercept

Regression Line Equation

In original units, the regression line has this equation:

$$\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}$$

estimated y in standard units

x in standard units

Lines can be expressed by *slope* & *intercept*

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

Regression Line Equation

In original units, the regression line has this equation:

$$\frac{\text{estimate of } y - \text{ average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x - \text{ average of } x}{\text{SD of } x}$$

what we want

what we observe

Lines can be expressed by *slope* & *intercept*

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

Slope and Intercept

estimate of y = slope * x + intercept

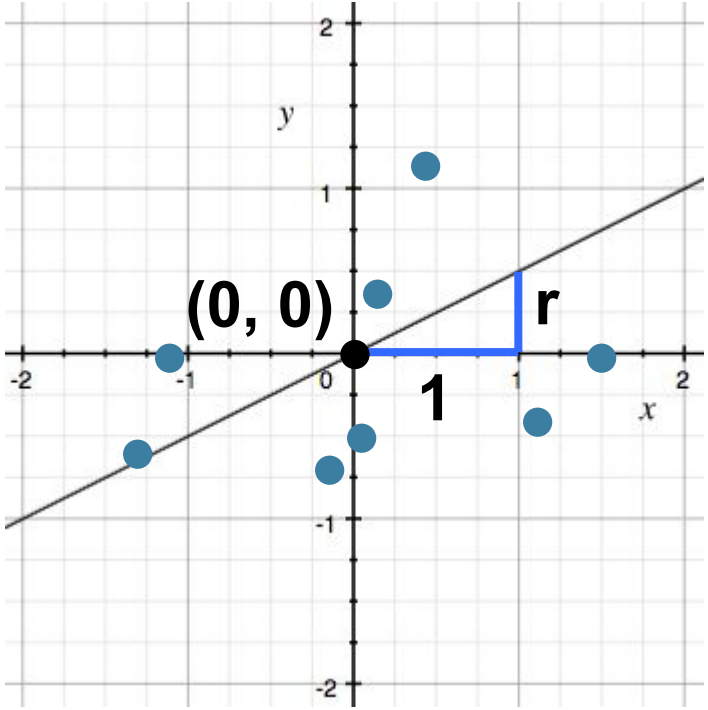
$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

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

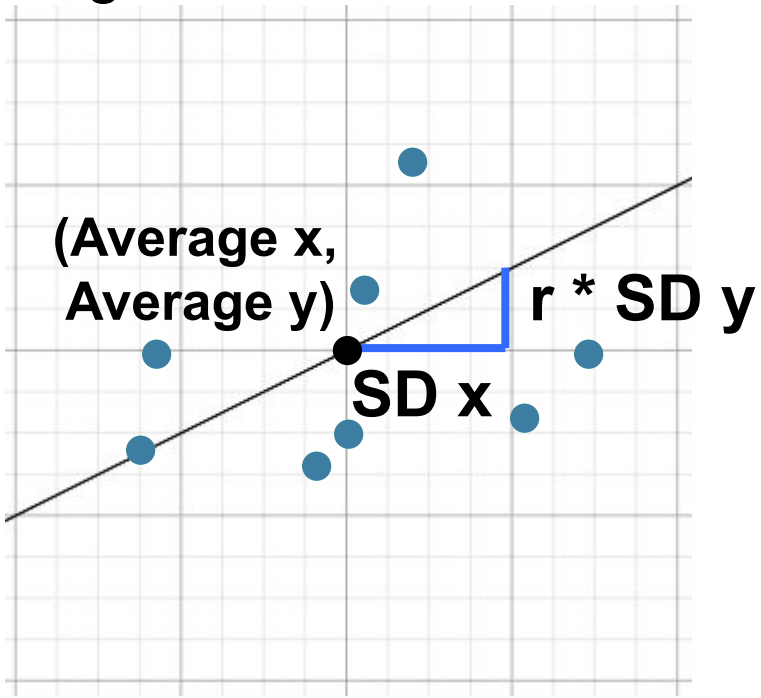
(Demo)

Regression Line

Standard Units



Original Units



Discussion Question

Suppose we use linear regression to predict candy prices (in dollars) from sugar content (in grams). What are the units of each of the following?

- r
- The slope
- The intercept

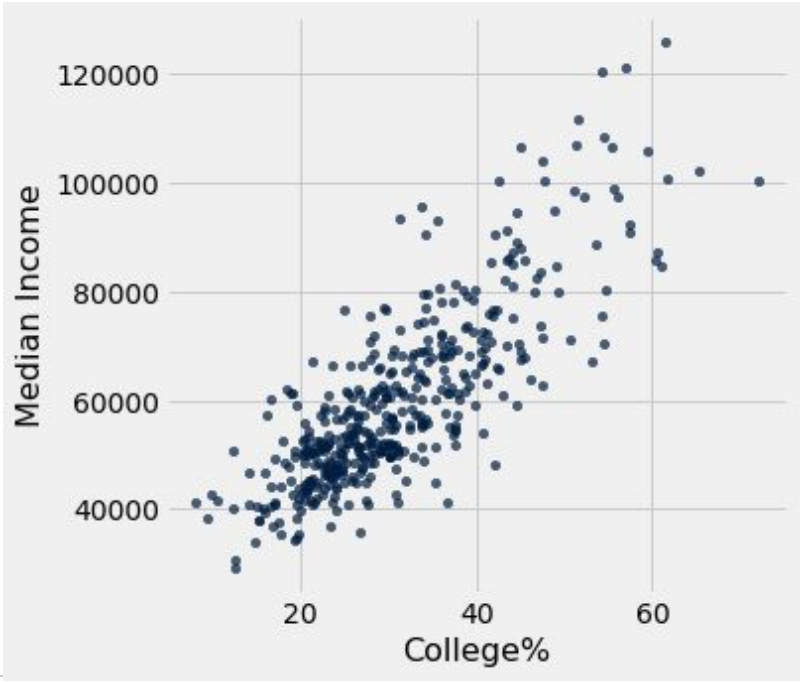
Least Squares

Discussion Question

Based only on the graph, which must be true? Explain.

1. Going to college causes people to get higher incomes.
2. For any district, having more college-educated people live there causes median incomes to rise.
3. For any district, having a higher median income causes more college-educated people to move there.

USA Congressional Districts, 2016



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

(Demo)

Numerical Optimization

- Numerical minimization is approximate but effective
- Lots of machine learning uses numerical minimization
- If the function `mse(a, b)` returns the mse of estimation using the line "estimate = a*x* + 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)