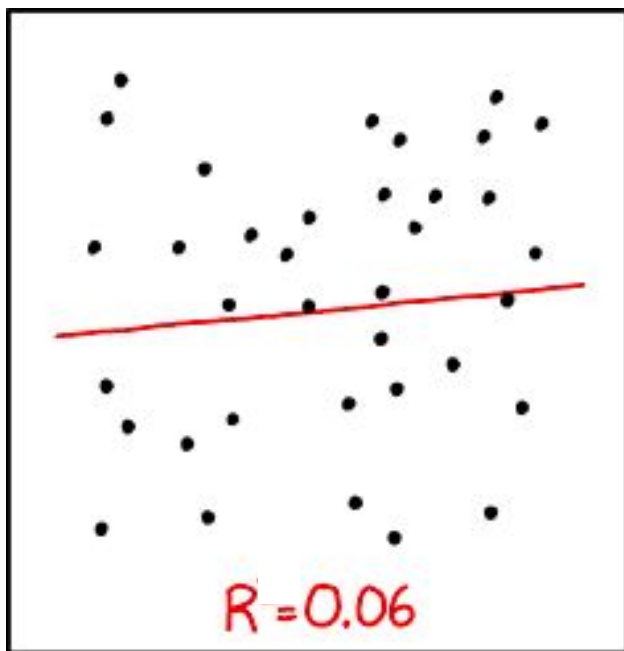




DATA 8
Spring 2022

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.

Announcements

- Project 2 checkpoint due tonight, 11:59pm PT
 - Entire project due Friday, 04/15
- Homework 8 and Lab 8 scores out today
- Homework 10 due Thursday, 04/14
 - Turn in on Wednesday for a bonus point

Regression Roadmap

- Monday
 - How to measure linear association
 - Wednesday
 - Predicting one numerical variable from a another
 - The regression line
 - **Today**
 - The “best” linear predictor
 - The method of least squares
-

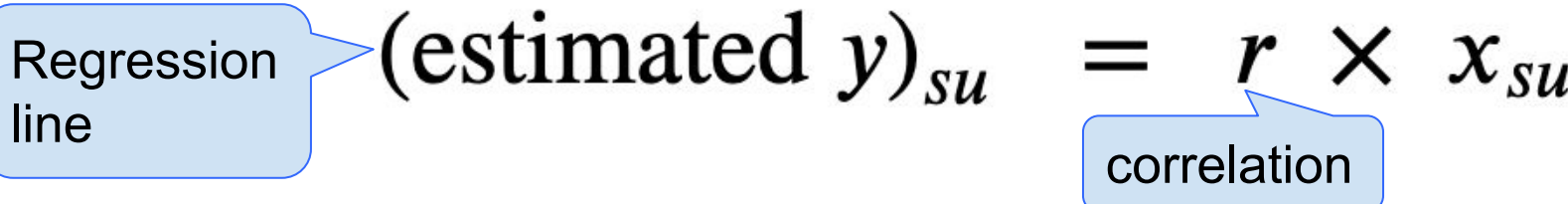
Linear Regression

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

$$(\text{estimated } y)_{su} = r \times x_{su}$$

correlation

Not true for all points — a statement about averages

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

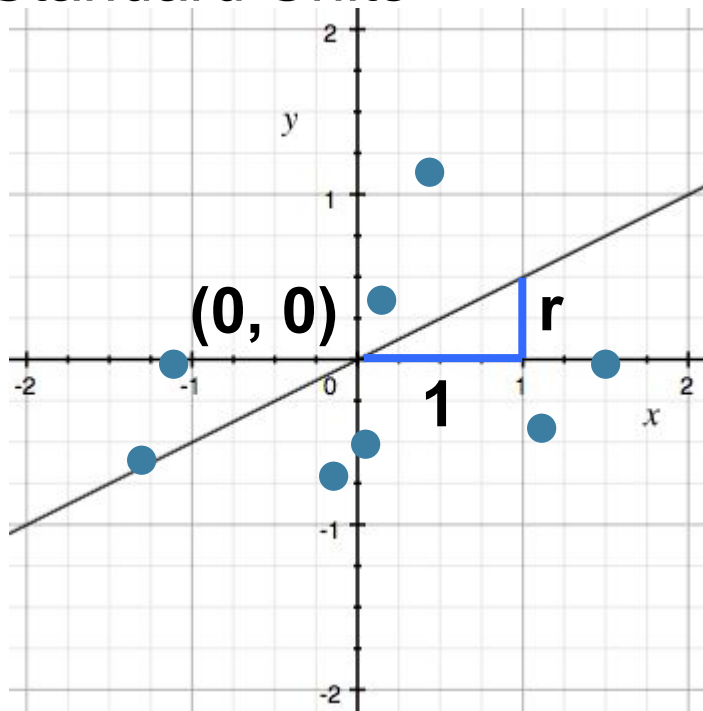
x in standard units

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

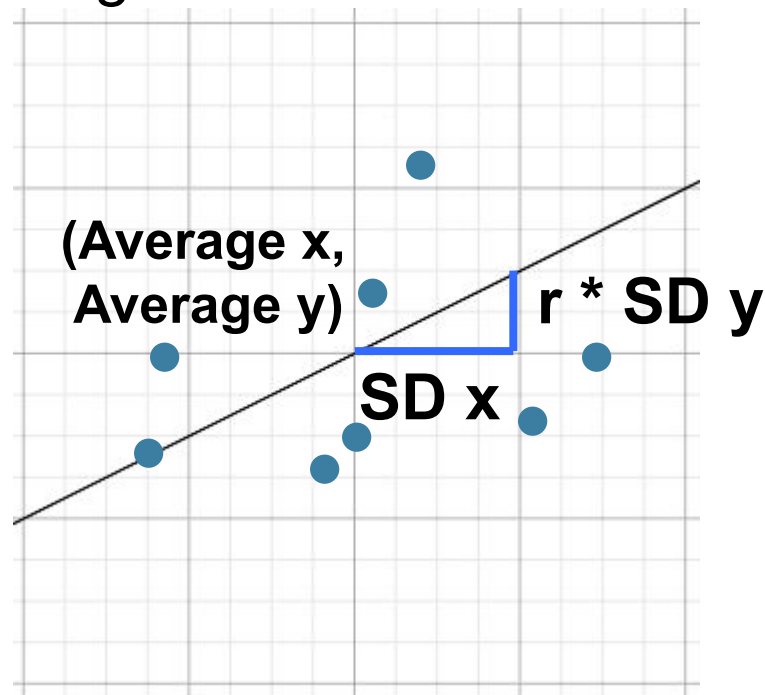
$$y = \text{slope} \times x + \text{intercept}$$

Regression Line

Standard Units



Original Units



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
-

Slope and Intercept (in original units)

estimate of y = slope * x + intercept

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

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

(Demo)

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)

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

(Demo)

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)

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 = $ax + b$ ”,
 - then `minimize(mse)` returns array `[a0, b0]`
 - `a0` is the slope and `b0` 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)

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
-