190F Foundations of Data Science

Spring 2020

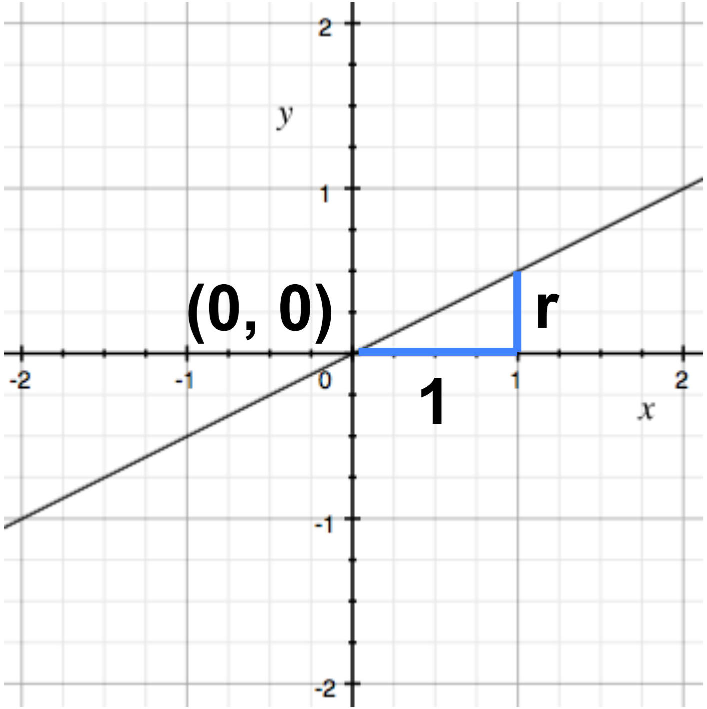
Lecture 18

Least Squares

Announcements

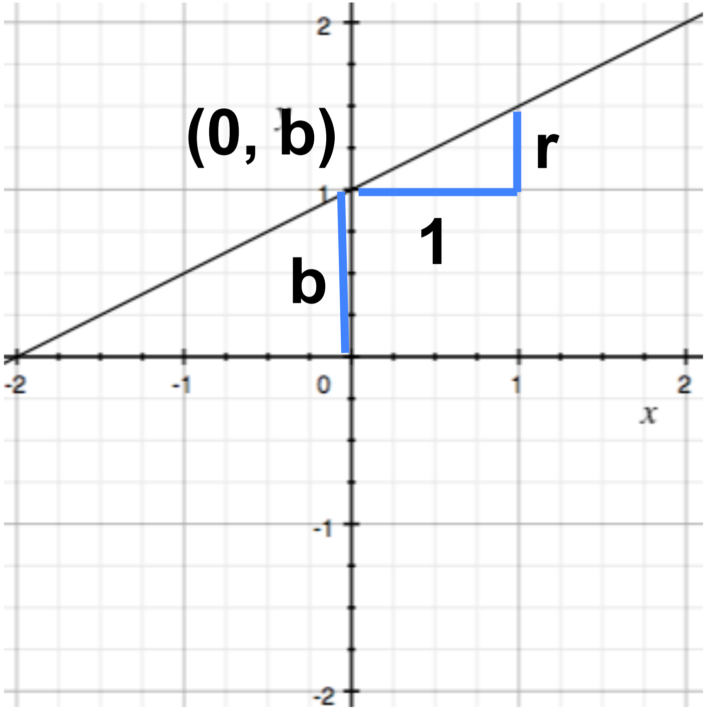
Algebra Review

Equation of a Line



$$y = r \times x$$

Equation of a Line

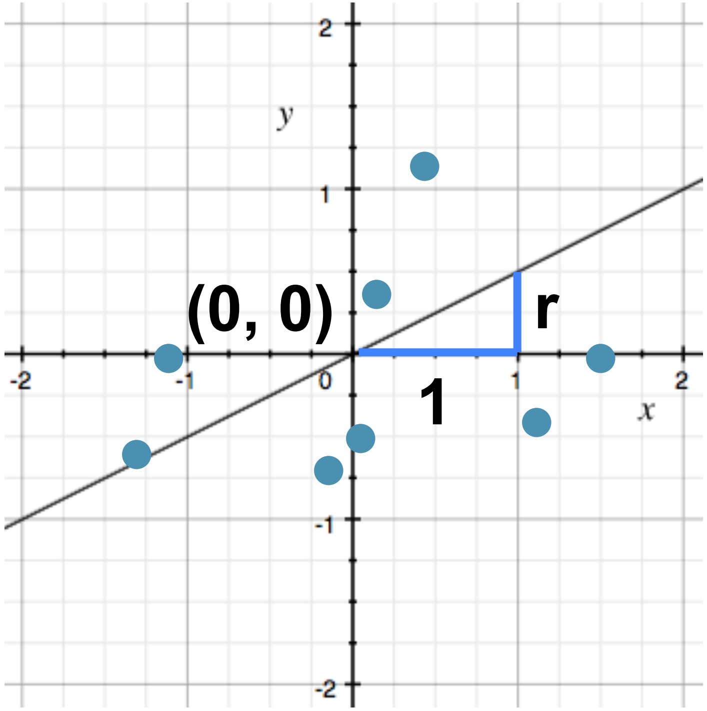


$$y = r \times x + b$$

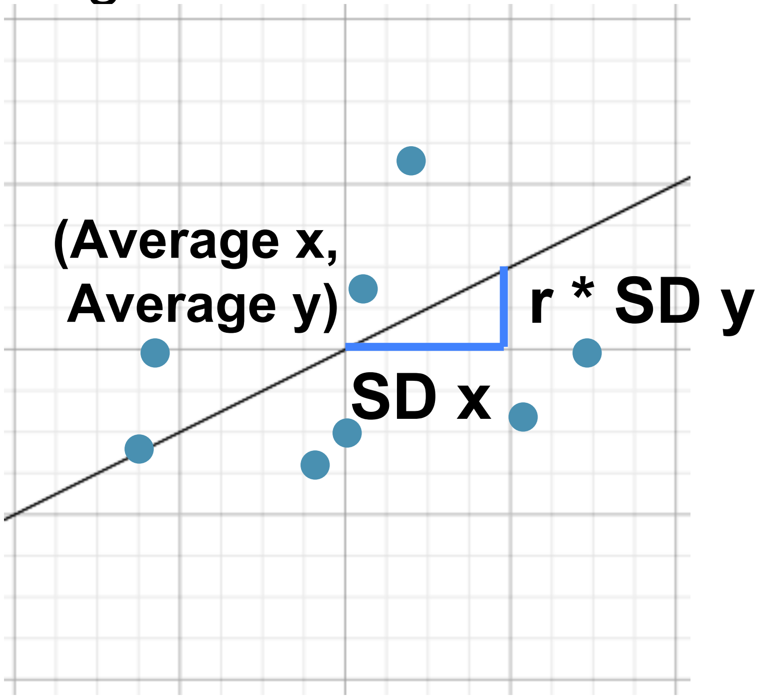
Linear Regression

Regression Line

Standard Units

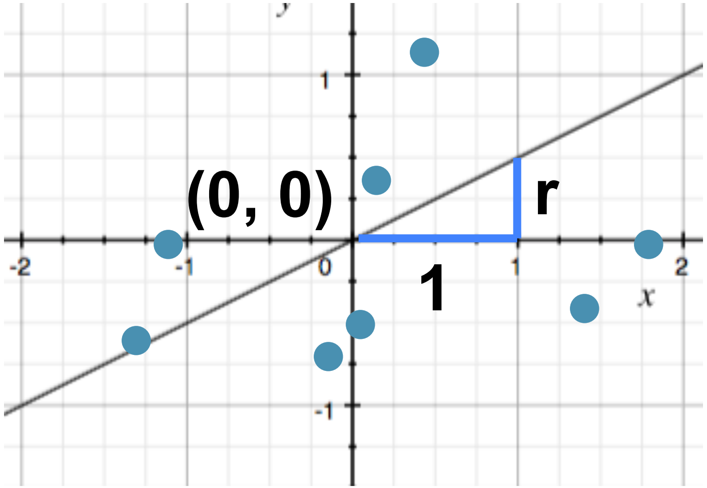


Original Units



Regression Line Equation

In standard units, the equation of the regression line is:



Fitted value

Observed value

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

Correlation coefficient

Regression Line Equation

In original units, the regression line has this equation:

$$\underbrace{\frac{\text{estimate of } y \; - \; \text{average of } y}{\text{SD of } y}}_{\text{y in standard units}} = r \times \underbrace{\frac{\text{the given } x \; - \; \text{average of } x}{\text{SD of } x}}_{\text{x in standard units}}$$

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

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

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

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?

y = 0.75 * 2 = 1.5

(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 erro**r (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 = 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)