DS2500
2/4 - Tues.

Admin

- HW3 due Fri 9pm
- exam NEXT Fri 2/14
  ↳ practice problems today
- 12:0 2/10 exam prep

Agenda

1. Correlation
2. Linear regression
3. Python

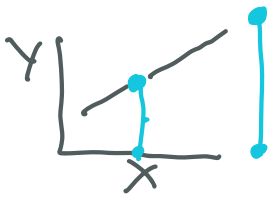
0. Today's Data Set
↳ northeastern data 2013-2024
- adm rate %
- selectiveness very, highly
- demographics raw #s
- tuition $

ethics/privacy
- source: new.edu
- accuracy ?
- consent
demographic concerns?

1. Correlation

↳ step one: understand relationship b/w 2 variables (ex: how correlated they are)

step two: if correlated, we can apply a prediction model (ex: linear regression)



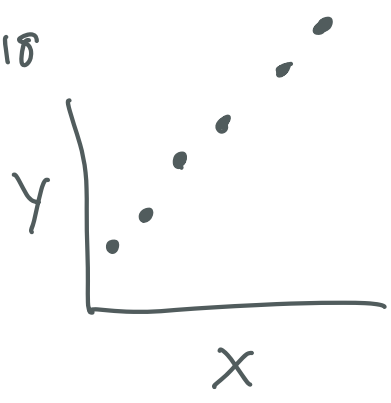
prediction model:
- fill in missing data
- predict future results

w/ correlation, we have:
- independent variable (x)
- dependent variable (y)

ex) X = 1, 2, 3, 4, 5, 6
Y = 8, 9, 13, 14, 17, 18

correlation
given a value of x, what is y?

prediction model
given a new value of x, what would y be?



- when x goes up, y goes up too
- when x goes up, y goes down

Correlation: r value

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r linear correlation

x_i, y_i indv values in dataset

$\bar{x}, \bar{y}$ mean

$X = 1,2,3$ $Y = 8,9,13$

$\bar{x} = 2$ $\bar{y} = 10$

Numerator

$(1-2)(8-10) + (2-2) + (3-2)(13-10)$

$2 + 0 + 3$

$\boxed{5}$

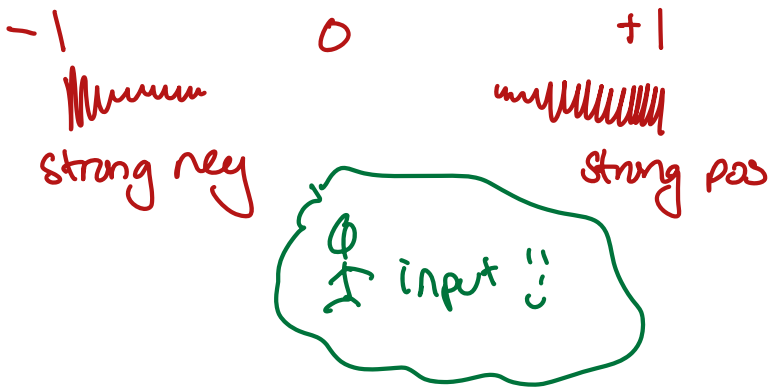
Denominator

$[(1-2)^2 + (0) + (3-2)^2] \cdot [(8-10)^2 + (9-10)^2 + (13-10)^2]$

$(1 + 0 + 1)(4 + 1 + 9)$

$\sqrt{28} = 5.29$

$r = 5/5.29 = .945$ very strong correlation!



In Python:

`statistics.correlation(x,y)`

If correlation is strong, then we can apply prediction model like linear regression

2. Linear regression

↳ prediction model - given a new x, what would y be?

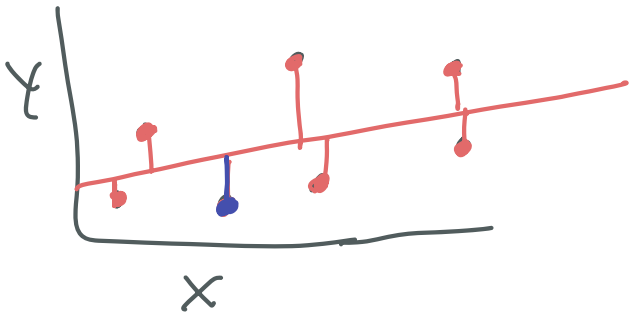
supervised learning - start with known x, y

linear regression: process of finding the line of best fit

line of best fit: minimizes mean Squared Error of all data points

↳ $y = mx + b$

slope intercept



- give Python X, Y values
- linear regression tries a bunch of lines and picks best one
- Best one = = smallest MSE where error is Euclidean distance →

(X) starting X, Y values

line of best fit $y' = mx+b$ $\quad y' = \frac{1}{2}x+3$

compute: Y', Error

X	Y	Y'	Error
2	6	4	2
4	2	5	3
6	9	6	3
8	2	7	5

Follow-up Qs

- predict a y for a new x
- why is error pretty easy?
- what is MSE?

→ $x=20$

$y = \frac{1}{2}20+3 = \boxed{13}$

→ euc $= \sqrt{(x-x')^2+(y-y')^2}$

$= \sqrt{(y-y')^2}$

$= |y-y'|$

$\frac{2^2+3^2+3^2+5^2}{4}$

$= \frac{4+9+9+25}{4} = \frac{47}{4} \approx \boxed{11.8}$

In Python:

- compute line of best fit: stats.linregress()
- draw line of best fit: sns.regplot()

4:24

3. Python

libraries:

- statistics (correlation)
- scipy (line equation)
- seaborn (draw line of best fit)

Goal:

↳ are adm rate and tuition correlated? If so, do a linear regression