§ 2.7 Empirical Models & Linear Regression

"Emperical" ~ based on observation

Suppose you observe $x_1, x_2, x_3, \dots, x_{N+1}$

- Population of a species after n generations $1 \le n \le N+1$
- Price of commodity after n days/weeks/months
- Proportion of a city's residents with college degree after $10n$ years
- Etc.

Q: Is there a $\boxed{\text{linear model}}$ that describes these observations?

$$x_1 \quad x_2 \quad x_3 \quad \dots \quad x_{N+1}$$

i.e. is there a pair of #'s $\boxed{a, b}$ such that

$$\boxed{x_{n+1} = a x_n + b} \quad \text{for } n = 1, 2, \dots, N$$

$\boxed{\text{If yes,}}$ then

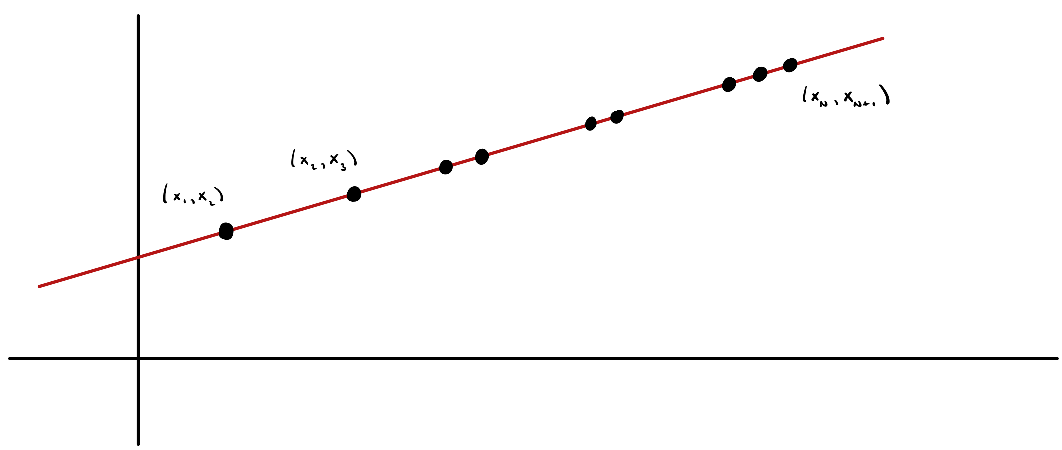
$$x_2 = a x_1 + b$$
$$x_3 = a x_2 + b$$
$$\vdots$$
$$x_{N+1} = a x_N + b$$

INPUT

OUTPUT

and the points $(x_1, x_2), (x_2, x_3), \dots, (x_N, x_{N+1})$

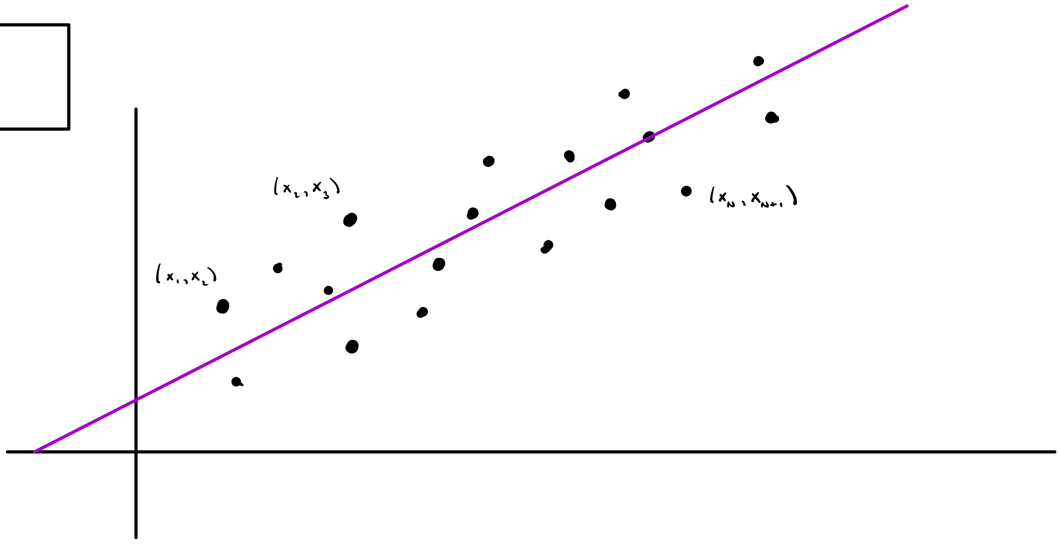
all lie on the graph $y = ax + b$



But Probably Not

That's OK.

Data probably not 100% accurate.



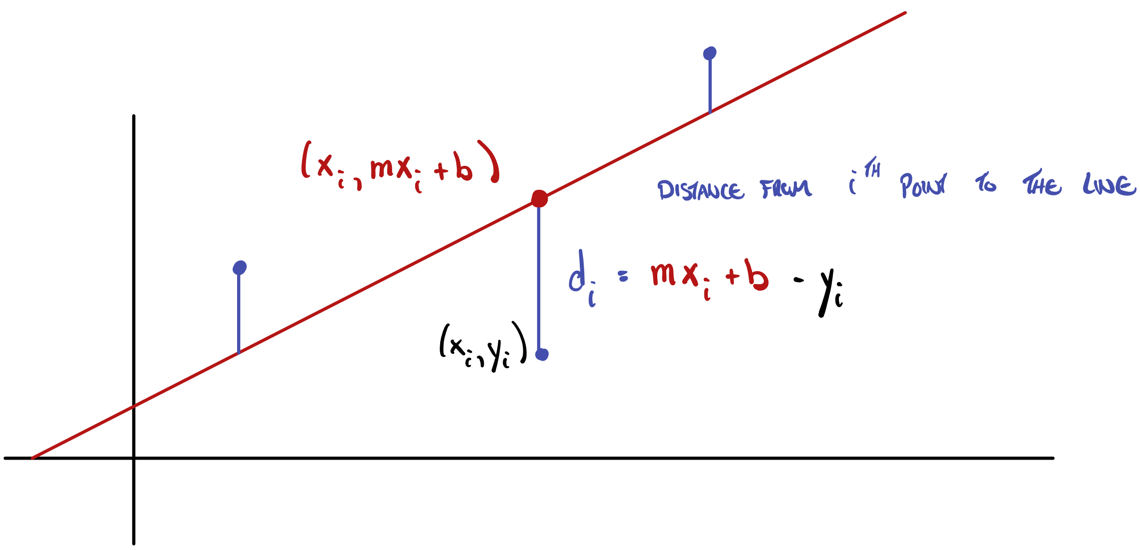
New Question: Can we find a pair of #'s a, b such that $x_{n+1} = a x_n + b$ is the linear model that best approximates the observations/data?

We should be more specific.

Let's pose this as a **Geometric Question:**

Given N points $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$.

Find m & b such that $y = mx + b$ is the unique straight line that minimizes the total error between the line & the points.



Total Error $E = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} (mx_i + b - y_i)^2$

$$E(m,b) = \sum_{i=1}^{N} (mx_i + b - y_i)^2$$

WE WANT TO FIND THE VALUES OF m & b THAT MINIMIZE THIS FUNCTION $E(m,b)$ OF 2 VARIABLES.

OPTIMIZATION

QUADRATIC POLYNOMIAL => DIFFERENTIABLE

> **FACT:** THE MINIMUM VALUE OF $E(m,b)$ WILL OCCUR AT THE POINT WHERE BOTH OF ITS PARTIAL DERIVATIVES ARE 0.
> (i.e. AT THE CRITICAL POINT OF E).

PARTIAL DERIVATIVES

GIVEN A FUNCTION OF 2 VARIABLES $f(x,y)$ IT HAS TWO PARTIAL DERIVATIVES.

- THE PARTIAL DERIV. OF f WRT x $\quad \frac{\partial}{\partial x} f = f_x$

  IS OBTAINED BY TAKING THE DERIV. OF f WRT x AND TREATING y (AND ANY FUNCTION OF y) AS A CONSTANT.

  e.g. $f(x,y) = 4x^2 - 3y^2 + 2x - 7y + 5$

  $f_x = 8x + 2$

- THE PARTIAL DERIV. OF f WRT y $\quad \frac{\partial}{\partial y} f = f_y$

  IS OBTAINED BY TAKING THE DERIV. OF f WRT y AND TREATING x (AND ANY FUNCTION OF x) AS A CONSTANT.

  e.g. $f(x,y) = 4x^2 - 3y^2 + 2x - 7y + 5$

  $f_y = -6y - 7$

ex. Let $f(x,y) = (7x-6y)^2 + 8x - y + 4$

FIND f_x & f_y.

ex. $f(x,y) = \sqrt{x^2+y^2} = (x^2+y^2)^{1/2}$

FIND f_x & f_y.

$$E(m,b) = \sum_{i=1}^{N} (mx_i + b - y_i)^2$$

m & b ARE VARIABLES

ALL x_i, y_i ARE CONSTANTS

MINIMIZED WHEN $E_m = E_b = 0$.

$$E_m = \sum_{i=1}^{N} 2(mx_i + b - y_i)x_i = 0$$

$$\sum_{i=1}^{N} (mx_i^2 + bx_i - x_iy_i) = 0$$

$$m\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_iy_i = 0$$

$$m\sum x_i^2 + b\sum x_i = \sum x_iy_i$$

$$E_b = \sum_{i=1}^{N} 2(mx_i + b - y_i) = 0$$

$$m\sum_{i=1}^{N} x_i + \sum_{i=1}^{N} b - \sum_{i=1}^{N} y_i = 0$$

$$m\sum x_i + Nb = \sum y_i$$

○ : CONSTANT (DETERMINED BY DATA)

WE MUST SOLVE THE SYSTEM OF m & b:

$$\boxed{\begin{aligned} \left(\sum x_i^2\right) m + \left(\sum x_i\right) b &= \sum x_iy_i \\ \left(\sum x_i\right) m + N b &= \sum y_i \end{aligned}}$$

NORMAL EQ'S FOR THE DATA

GENERALLY, THIS SYSTEM WILL HAVE A UNIQUE SOL'N (m,b).

THE LINE $y = mx + b$ WITH THESE VALUES BEST APPROXIMATES THE DATA $(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$.

THIS LINE IS CALLED THE LEAST-SQUARES REGRESSION LINE

ex. *In Exercises 17–20 first write the Normal Equations using the given data, and then solve to find the least squares line* y = mx + b.

* 17. (0, 0), (1, 2), (2, 3)
18. (0, 0), (0, 1), (1, 0), (1, 1)
19. (−2, 10), (−1, 6), (0, 5), (0, 4), (2, 0)
20. (−2, −5), (−1, −2), (0, 0), (1, 1), (2, 1)

$$\boxed{\begin{aligned} \left(\sum x_i^2\right) m + \left(\sum x_i\right) b &= \sum x_i y_i \\ \left(\sum x_i\right) m + N b &= \sum y_i \end{aligned}}$$

i	x_i	y_i	x_i^2	$x_i y_i$
1	0	0	0	0
2	1	2	1	2
$N = 3$	2	3	4	6
	3	5	5	8

$\sum x_i = 3$ $\qquad$ $\sum x_i^2 = 5$

$\sum y_i = 5$ $\qquad$ $\sum x_i y_i = 8$

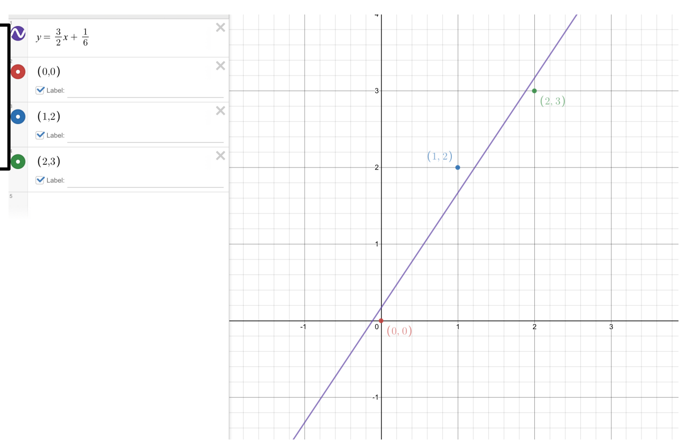
$$\begin{aligned} 5m + 3b &= 8 \\ -\left(\; 3m + 3b = 5 \;\right) & \\ \hline 2m &= 3 \quad \Rightarrow \quad m = \frac{3}{2} \end{aligned}$$

$$5\left(\frac{3}{2}\right) + 3b = 8 \quad \Rightarrow \quad b = \frac{8 - \frac{15}{2}}{3} = \frac{1}{6}$$

$m = 3/2$

$b = 1/6$

$$\boxed{y = \frac{3}{2}x + \frac{1}{6}}$$



https://www.desmos.com/calculator/lllkqqounf

In Exercises 21–24 use the given data to construct the best approximate linear iterative model.

21. $x_0 = 1, x_1 = 3, x_2 = 10, x_3 = 20$

* 22. $S_0 = 2, S_1 = 5, S_2 = 10, S_3 = 30$

23. $P_0 = 60, P_1 = 25, P_2 = 10, P_3 = 5, P_4 = 2, P_5 = 1$

24. $T_0 = 50, T_1 = 20, T_2 = 12, T_3 = 8, T_4 = 5, T_5 = 4$

25. (a) Instead of using the Chain Rule, find the partial derivatives E_m and E_b of $E = \sum_{i=1}^{N}(mx_i + b - y_i)^2$ by first squaring each term being added and then differentiating. (b) Show that this yields the same E_m and E_b as before.

$x_{n+1} = a x_n + b$

$y = mx + b$

(2,5) (5,10) (10,30)

$2 \rightarrow 5 \rightarrow 10 \rightarrow 30$

INPUT x OUTPUT y

https://www.desmos.com/calculator/ryxeOmpuvd

	i	x_i	y_i	x_i^2	$x_i y_i$
	1	2	5	4	10
	2	5	10	25	50
$N =$	3	10	30	100	300
		17	45	129	360

$$\left(\Sigma x_i^2\right) m + \left(\Sigma x_i\right) b = \Sigma x_i y_i$$
$$\left(\Sigma x_i\right) m + N b = \Sigma y_i$$

NORMAL EQ'S:

$$\left.\begin{aligned} 129\,m + 17\,b &= 360 \\ 17\,m + 3\,b &= 45 \end{aligned}\right\} \text{ SOLVE FOR } m, b$$

$$y = \frac{315}{98}x - \frac{315}{98}$$

$$\boxed{x_{n+1} = \frac{315}{98}x_n - \frac{315}{98}}$$