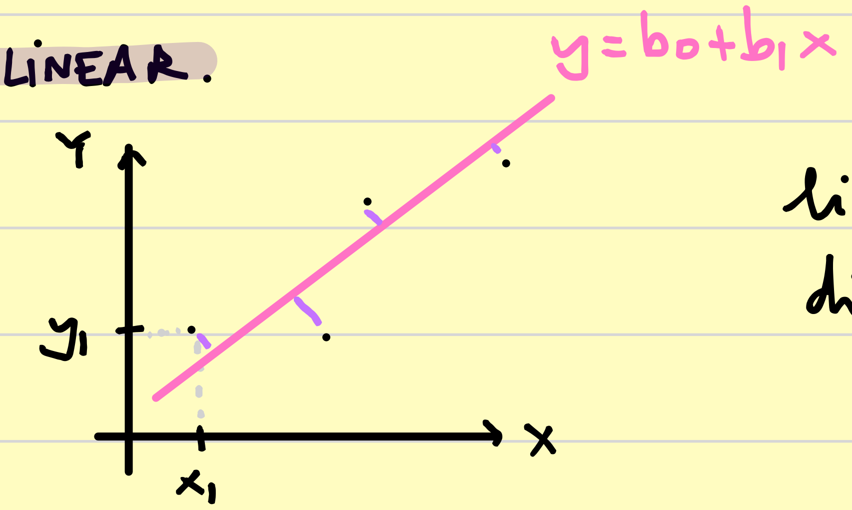
Regression Algorythms. Predictions based on data.

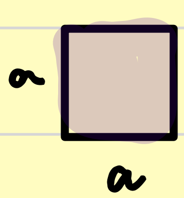
- LINEAR
- POLYNOMIC
- LOGISTIC

Hypothesis. we have gathered data from a System.
Goal. Predict the behaviour of the System based on the data.

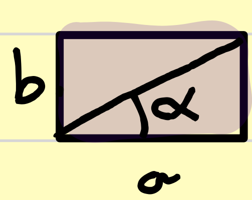
LINEAR.



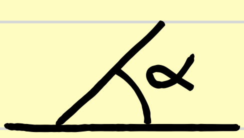
Goal. find the equation of the line $y=b_0+b_1x$ that has minimum distance to the points of the dataset.



Area $= a^2$



Area $= a \cdot b$



$tg\alpha = \frac{b}{a}$

Step 1. Gather data ✓

	x	y
cw1	3	6'5
cw2	4	8'5
cw3	6	13
cw4	3	3'5

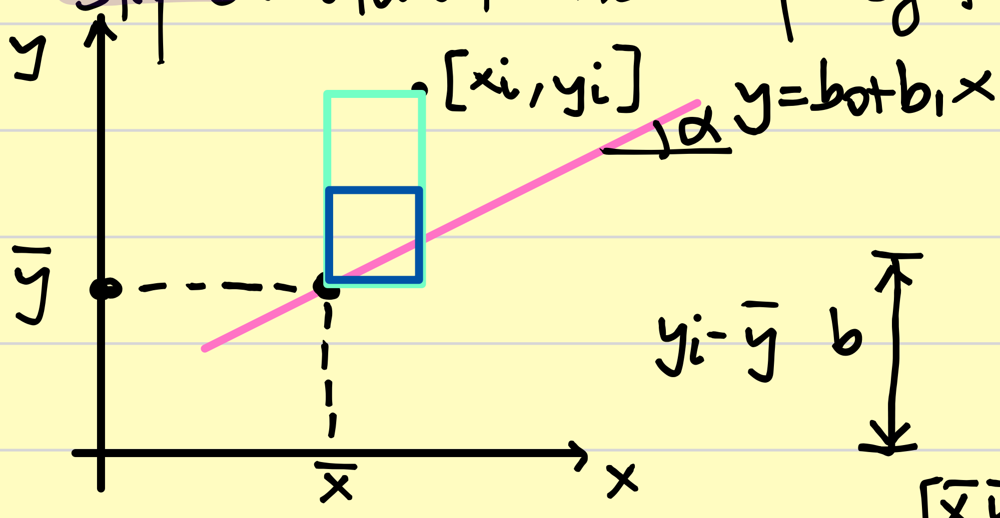
Step2. Mean value of the variables

** The linear regression crosses the mean value: $\bar{y}=b_0+b_1\bar{x}$ **

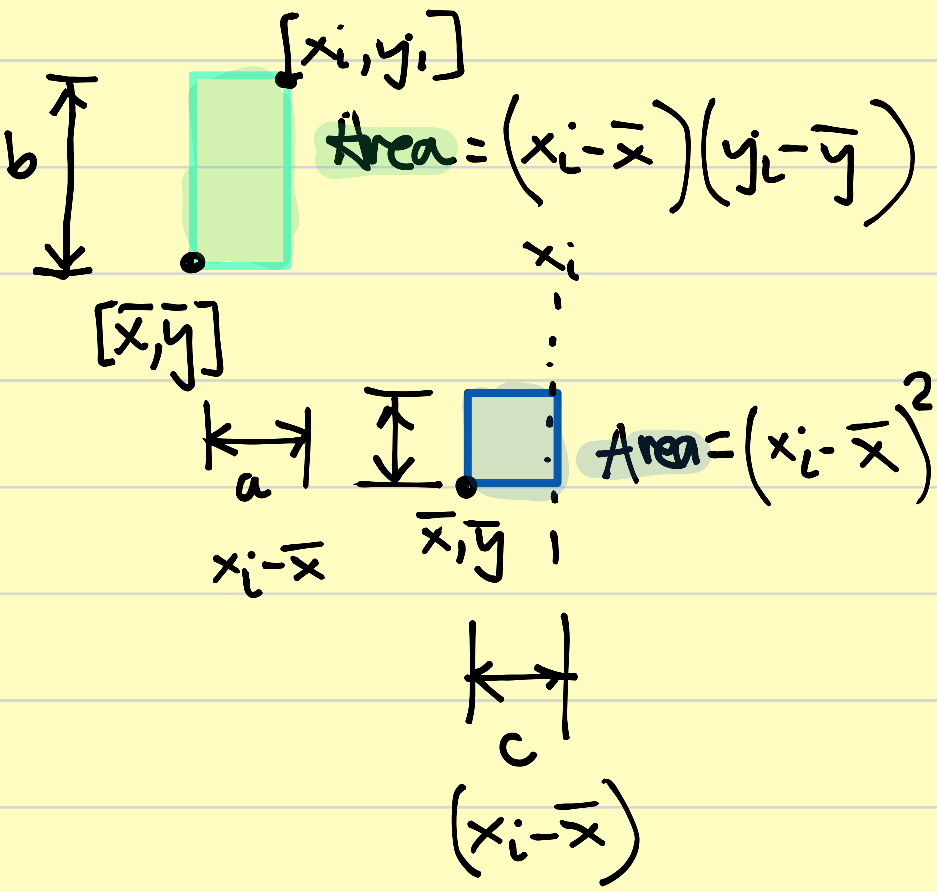
$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i=\frac{1}{4}\left[3+4+6+3\right]=4$$

$$\bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i=\frac{1}{4}\left[6'5+8'5+13+3'5\right]=7'87$$

Step 3. Calculate the slope of the line $[b_1]$



$tg\alpha = b_1$



$$\boxed{b_1 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}}$$

$$b_1 = \frac{(3-4)(6'5-7'87)+(4-4)(8'5-7'87)+(6-4)(13-7'87)+(3-4)(3'5-7'87)}{(3-4)^2+(4-4)^2+(6-4)^2+(3-4)^2}$$

$$= \frac{29}{6} = 4'83 \rightarrow y = b_0 + 4'83\,x$$

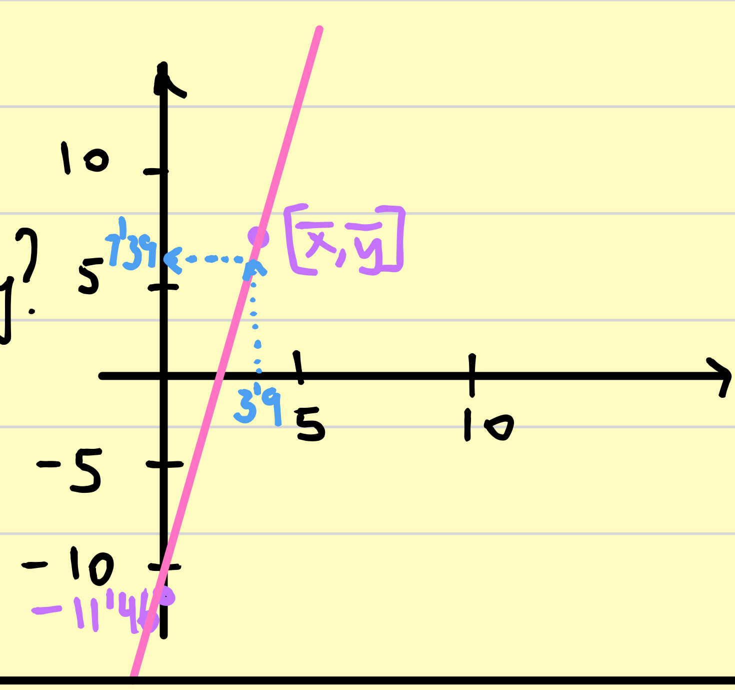
Step 4. The line goes through $[\bar{x}, \bar{y}]$

$$\bar{y} = b_0 + 4'83 \cdot \bar{x} \longrightarrow 7'87 = b_0 + 4'83 \cdot 4 \rightarrow b_0 = -11'46$$

$$\boxed{y = -11'46 + 4'83\,x}$$

If $x = 3'9$, what is the predicted value of y?

$$y_{[x=3'9]} = -11'46 + 4'83 \cdot 3'9 = 7'39$$



POLYNOMIC (NON LINEAR)

Goal. We aim to find out the relationship btw. a variable x and other variable Y with help of a non-linear regression.

Step 1. Gather data ✓

	x	y
cw_1	1	2'5
cw_2	2	5'8
cw_3	3	11'9
cw_4	4	21'4
cw_5	5	31'2

Step 2. Hypothesis: the regression is of order 2 (wadratic)

$$y = a x^2 + b x + c$$

1. $[1, 2'5] \longrightarrow 2'5 = a \cdot 1^2 + b \cdot 1 + c \quad (1)$
2. $[2, 5'8] \longrightarrow 5'8 = a \cdot 2^2 + b \cdot 2 + c \quad (2)$
3. $[3, 11'9] \longrightarrow 11'9 = a \cdot 3^2 + b \cdot 3 + c \quad (3)$

$(2)-(1) \rightarrow 5'8 - 2'5 = 4a - a + 2b - b + c - c = 3a + b$

$3'3 = 3a + b \quad (4)$

$(3)-(2) \longrightarrow 11'9 - 5'8 = 9a - 4a + 3b - 2b + c - c = 5a + b$

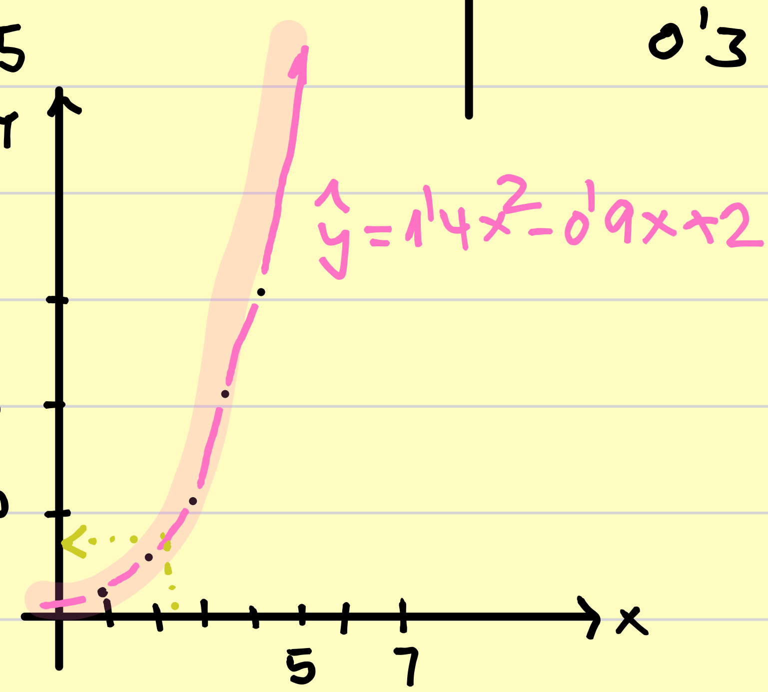
$6'1 = 5a + b \quad (5)$

$(5)-(4) \longrightarrow 6'1 - 3'3 = 5a - 3a + b - b = 2a$

$2'8 = 2a \rightarrow a = 1'4 \xrightarrow[(5)]{} b = -0'9 \xrightarrow[(1)]{} c = 2$

$$\boxed{\hat{y} = 1'4x^2 - 0'9x + 2}$$

	x	y	Prediction $\hat{y} = 1'4x^2 - 0'9x + 2$	Error $(\hat{y} - y)$
Cw1	1	2'5	$1'4(1^2) - 0'9 \cdot 1 + 2 = 2'5$	0
Cw2	2	5'8	$1'4(2^2) - 0'9 \cdot 2 + 2 = 5'7$	-0'1
Cw3	3	11'9	11'9	0
Cw4	4	21'4	21'4	0
Cw5	5	31'2	31'5	0'3



What is the value of y in CW7?

$$\hat{y}_{[CW_7]} = 1'4 \cdot 7^2 - 0'9 \cdot 7 + 2 = 64'3$$

And in CW 2'5?

$$\hat{y}_{[CW_{2'5}]} = 1'4 \cdot 2'5^2 - 0'9 \cdot 2'5 + 2 = 8'5$$

LOGISTIC.

LR is a type of regression used when the dependent variable (y) is binary (e.g. 0·1, Yes·No, Pass·Fail). LR predicts the probability of the outcome based on one or more independent variables (x_i).

Unlike linear/polynomic regression, where we predict a continuous value, logistic regression outputs probabilities that are mapped to a binary decision using a threshold (e.g. $p > 0'5$).

Example:

A professor wants to predict whether a student will Pass ($y=1$) or Fail ($y=0$) an exam based on the number of hours they study (x).

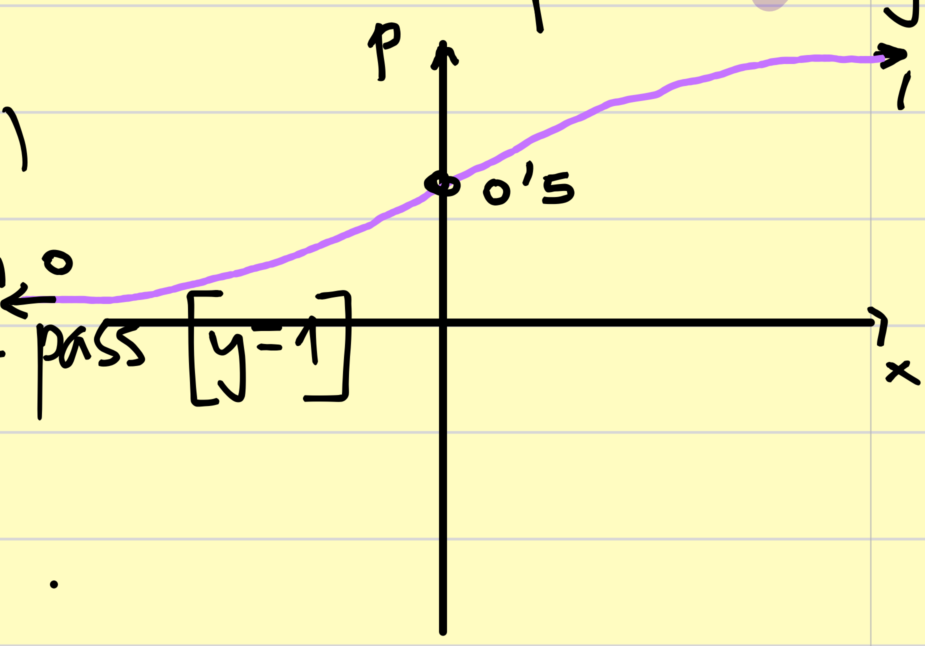
Step 1. Gather data ✓

x [hours]	y [Pass [1], Fail [0]]
1	0
2	0
3	0
4	1
5	1
6	1

Step 2. LR Model.

LR uses the logit function to model the relationship btw x and y

$$\boxed{\ln\left[\frac{p}{1-p}\right] = b_0 + b_1 x} \quad (1)$$



where: p is the probability of pass [$y=1$]

b_0 is the intercept

b_1 is the slope of x.

$$\frac{p}{1-p} = e^{b_0+b_1x} \rightarrow p = e^{b_0+b_1x} - p\left[e^{b_0+b_1x}\right] \rightarrow$$

$$\rightarrow p\left[1+e^{b_0+b_1x}\right] = e^{b_0+b_1x} \rightarrow \boxed{p = \frac{1}{1+e^{-(b_0+b_1x)}}}$$

Step 3. Solve

For simplicity, we will estimate the coefficients

$x = 3 \ , \ y = 0$

$x = 4 \ , \ y = 1$

** We assume $x = 4 \rightarrow p = 0'5 \xrightarrow{(1)} \ln\left[\frac{0'5}{1-0'5}\right] = b_0 + b_1 \cdot 4$

First we take the point x=4 where transition starts: p=0'5

** We assume $x = 6 \rightarrow p = 0'9 \xrightarrow{(1)} \ln\left[\frac{0'9}{1-0'9}\right] = b_0 + b_1 \cdot 6$

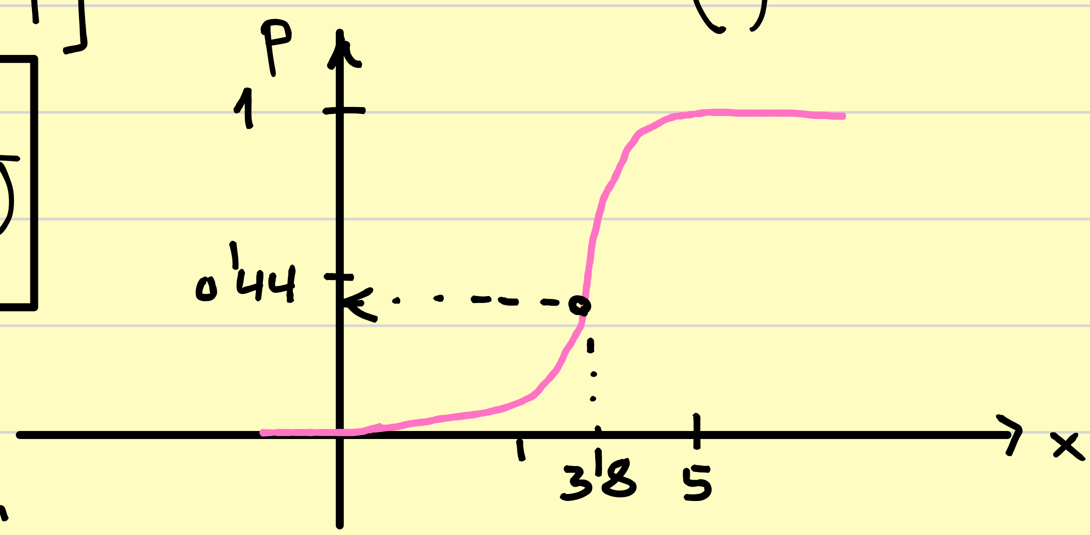
Second we take the point x=6 where transition ends: p=0'9.

$$\ln[1] = 0 = b_0 + b_1 4 \qquad (2)$$

$$\ln\left[\frac{0'9}{0'1}\right] = b_0 + b_1 \cdot 6 \qquad (3)$$

$$(3)-(2) \rightarrow \ln\left[\frac{0'9}{0'1}\right] = 2b_1 \rightarrow b_1 = 1'1 \xrightarrow[(2)]{} \ , = -4'4$$

$$\boxed{p = \frac{1}{1 + e^{-(-4'4 + 1'1x)}}}$$



• What is the probability of a student passing the exam after studying 3'8 hours?

$$p = \frac{1}{1 + e^{-(-4'4 + 1'1 \cdot 3'8)}} = 0'44$$