Régression Linéaire

Contexte : Relation entre

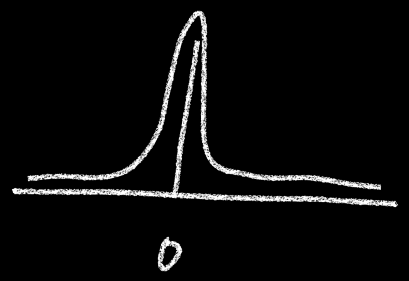
- une variable numerique target y

- une ou plusieurs variables explicatives $X_1, \dots, X_p$

Hypothèse d'une relation linéaire.
Caractéristiques de cette rel. Validité ?

Relation linéaire entre y et $x_1, \dots, x_p$:

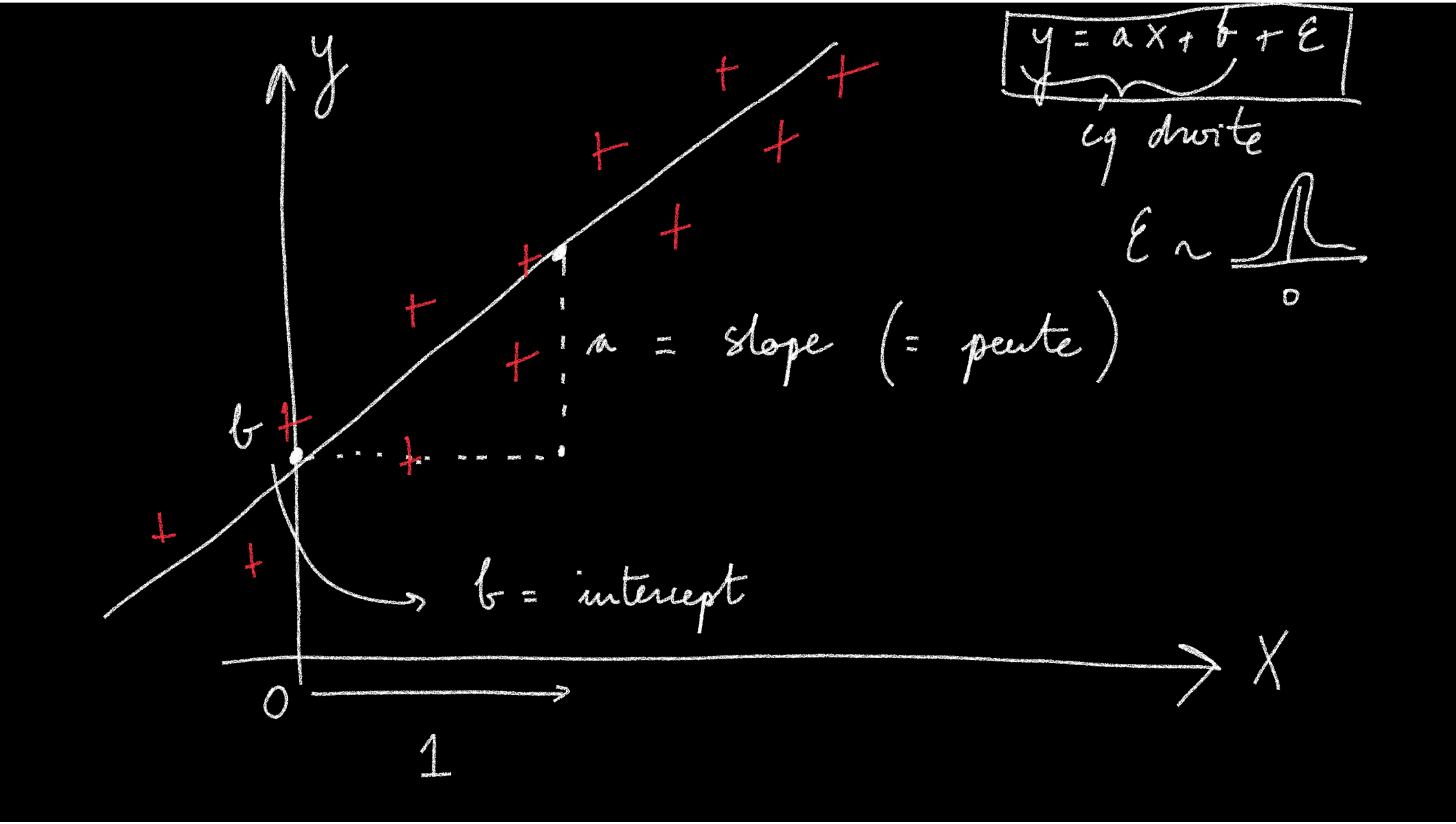
$$y = \alpha_1 \cdot x_1 + \alpha_2 \cdot x_2 + \dots + \alpha_p \cdot x_p + \alpha_{p+1} + \varepsilon$$

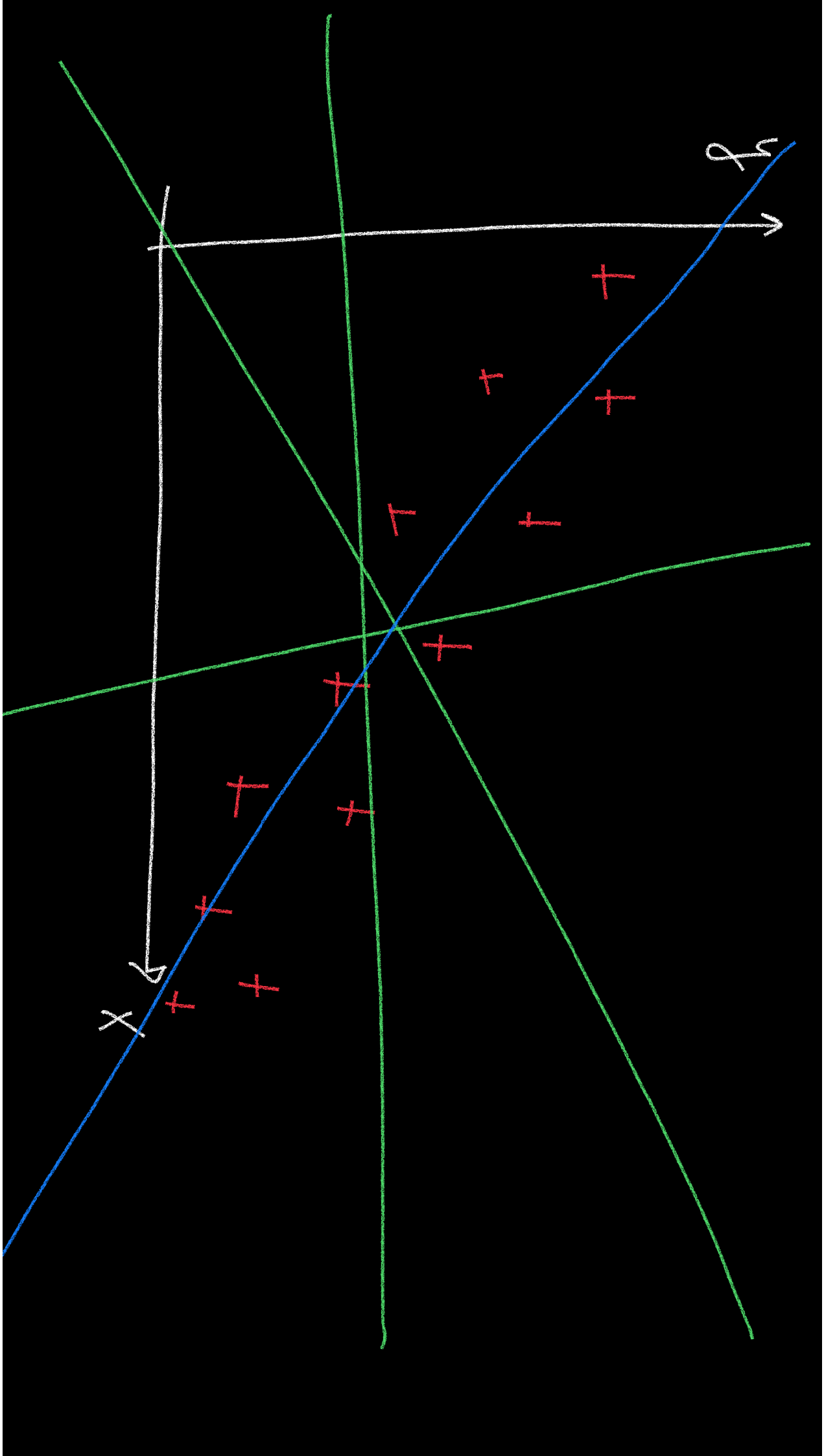
$\varepsilon \overset{\sim}{\uparrow}$ 

Suit une loi

$y = \alpha_1 x_1 + \alpha_2 + \varepsilon$

équation d'une droite


$y = aX + b + \varepsilon$
eq droite
$\varepsilon \sim$
0
y
a = slope (= pente)
b
b = intercept
0
1
X


q5
y
x

Régression Linéaire avec Moindres Carrés

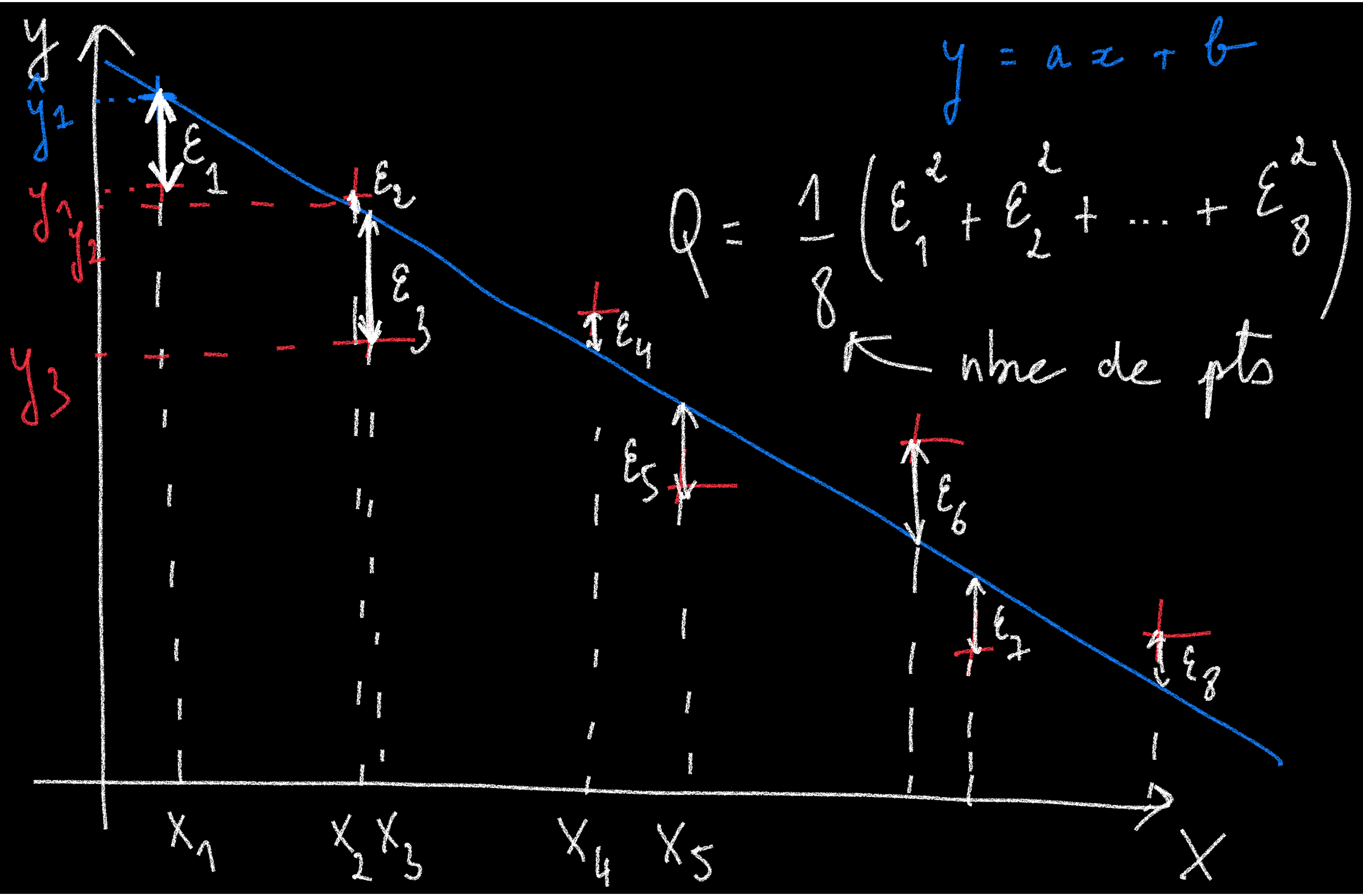
cherche la droite qui minimise

$$\frac{1}{N} \sum_{i=1}^{N} \text{erreur}^2$$

N : nbre de pts

$$\text{erreur}[i] = \underbrace{y[i]}_{\text{réel}} - \underbrace{(a\,X[i] + b)}_{\text{prédiction}}$$

$\hat{y}[i]$



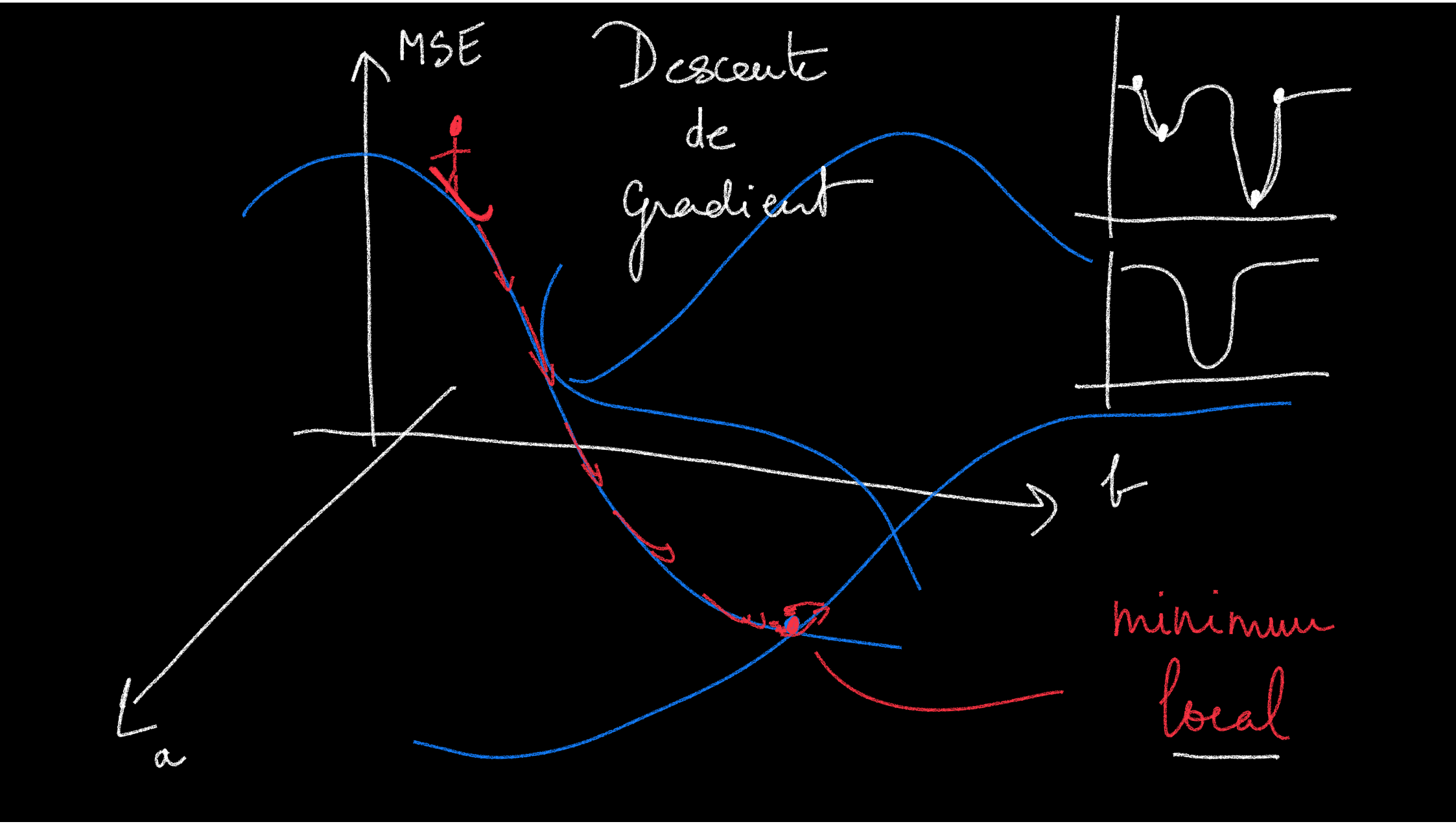
y = ax + b
Q = 1/8 (ε1² + ε2² + ... + ε8²)
nbre de pts
y
y1
y2
y3
ε1
ε2
ε3
ε4
ε5
ε6
ε7
ε8
X1
X2
X3
X4
X5
X

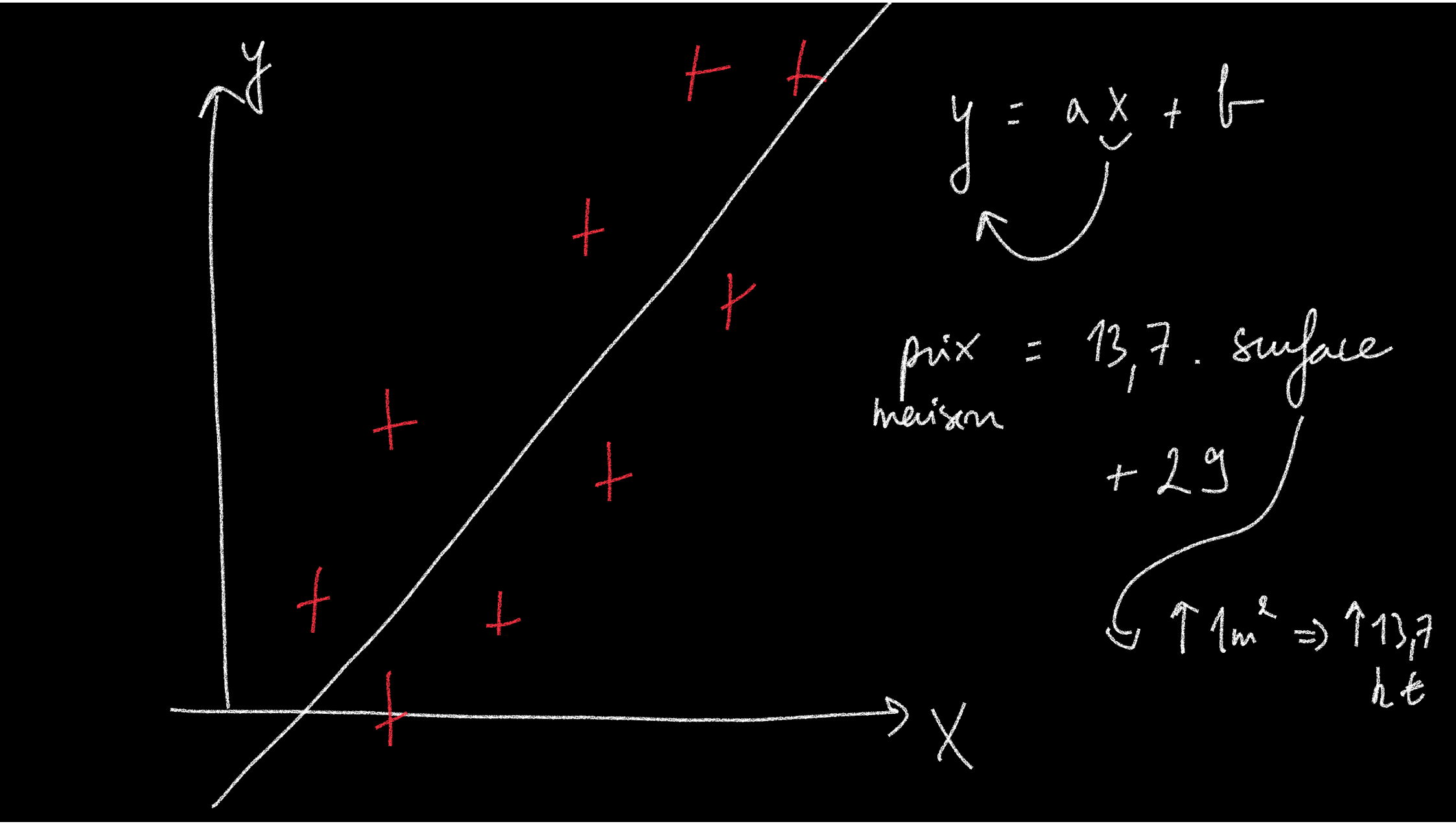
Mean Square Error $= \frac{1}{N} \sum_{i=1}^{N} \underbrace{(y[i] - \hat{y}[i])^2}_{\geq 0} \Big\} \geq 0$

où $\hat{y}$ = prédictions

y = réel

Rég Lin avec Moindres Carrés : $\underset{a,b}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} (y[i] - (a x[i] + b))^2$


MSE
Descente de Gradient
b
a
minimum local


y
x
y = ax + b
prix maison = 13,7 . surface + 29
↑ 1m² ⇒ ↑ 13,7 k€

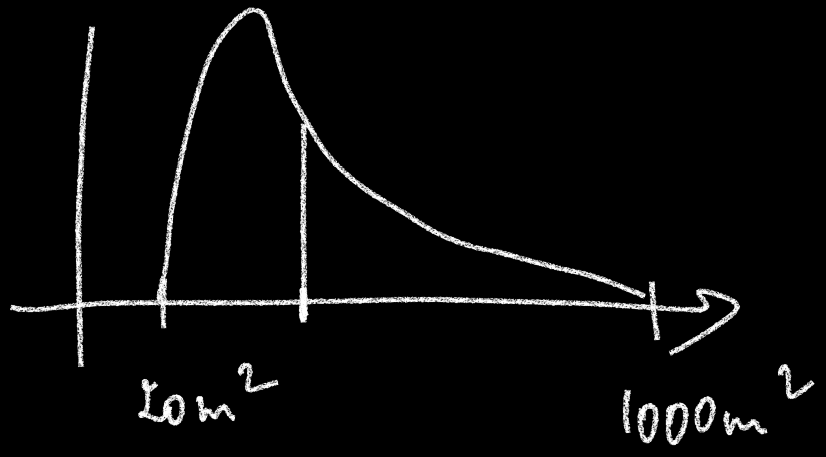
$$y = \alpha_1 x_1 + \dots + \alpha_p x_p + \alpha_{p+1}$$

prix maison $= 13{,}7 \cdot$ surface
$+ 3{,}8 \cdot \#$ pièces
$- 2{,}8 \cdot \#$ decibels
$+ \dots$

surface $> \#$ pièces

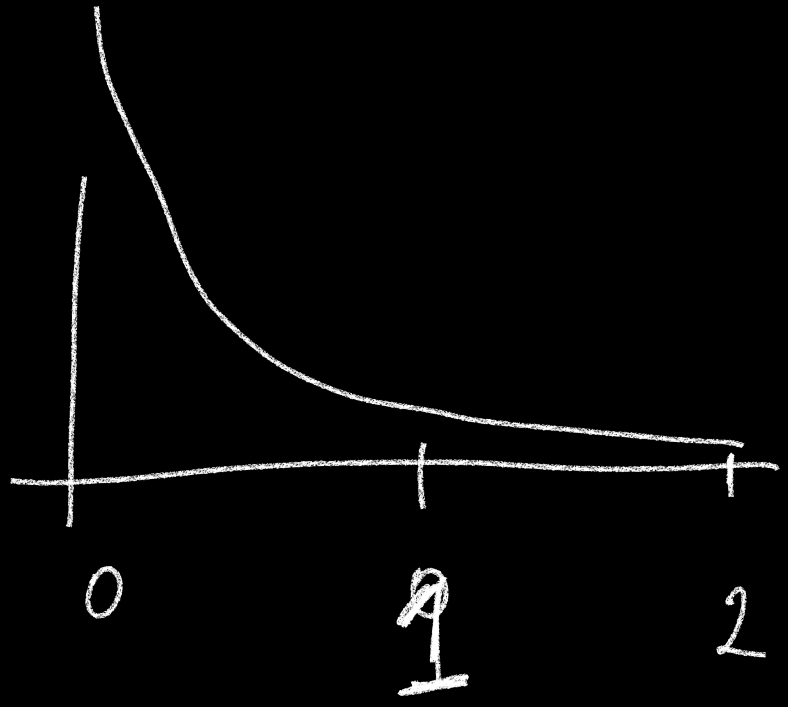
$\#$ decibels

Surfaces

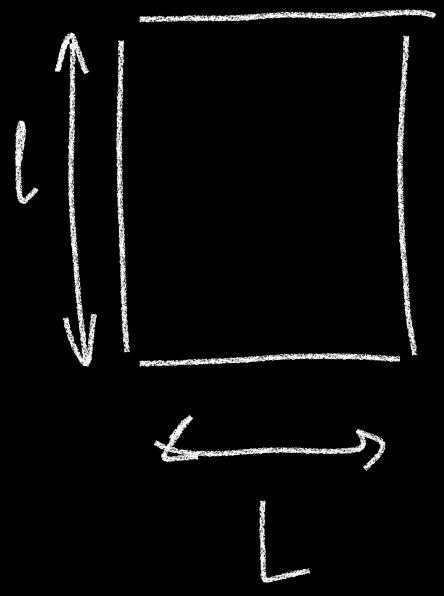


écart-type
$\sim 200\ m^2$

garages



écart-type
$\sim 0,15$



l

10000

10000000

L

1

1000

prix maison

=

37 . (l) cm

+

2400 . (L) m

+

...

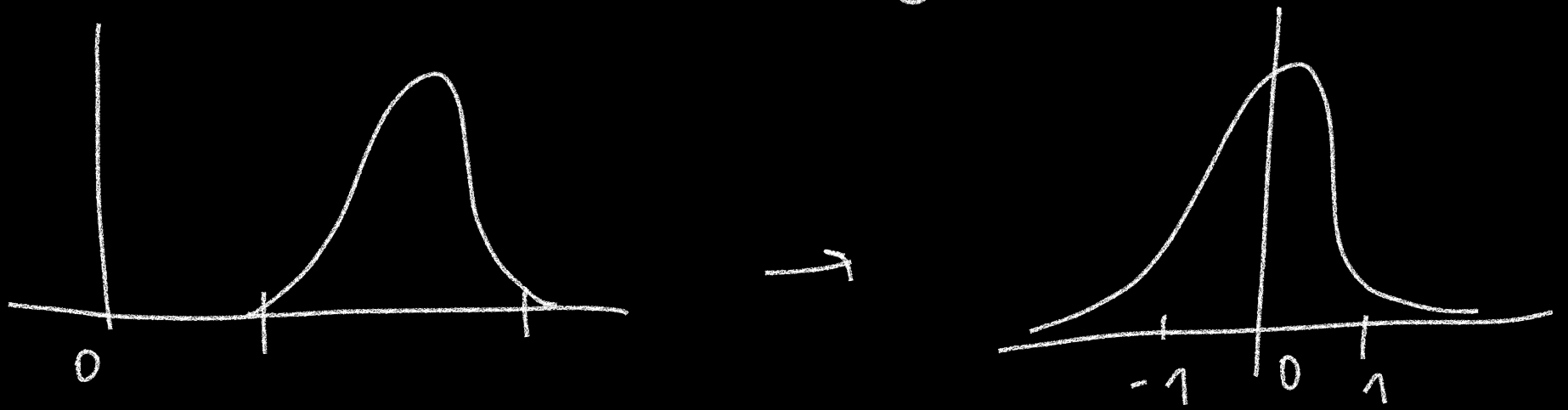
→ 3700 € à chaque mètre

→ 2400 €

Standardisation :

moyenne 0
écart-type 1

$$L \leftarrow \frac{L - \mu}{\sigma}$$



ℓ (cm)	L (m)
3000	42
10 000	53
100 000	17
57 000	24

$$prix = 37 \cdot \ell + 2400 \cdot L$$

$\ell \uparrow 1 \Rightarrow prix \uparrow 37$

$L \uparrow 1 \Rightarrow prix \uparrow 2400$

effet L > effet ℓ

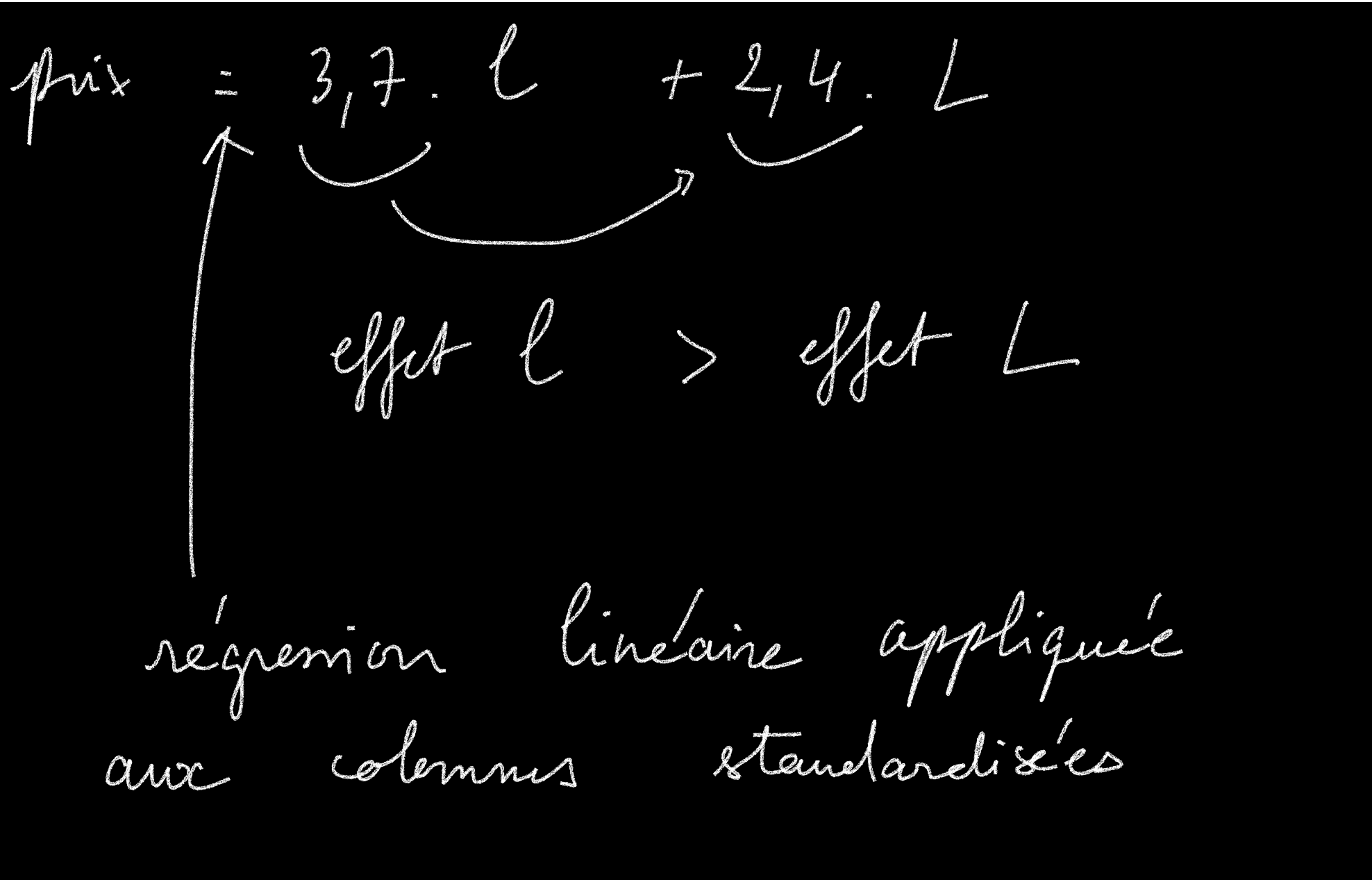
	effet ℓ	effet L
μ	42 500	34
σ	39 157	14

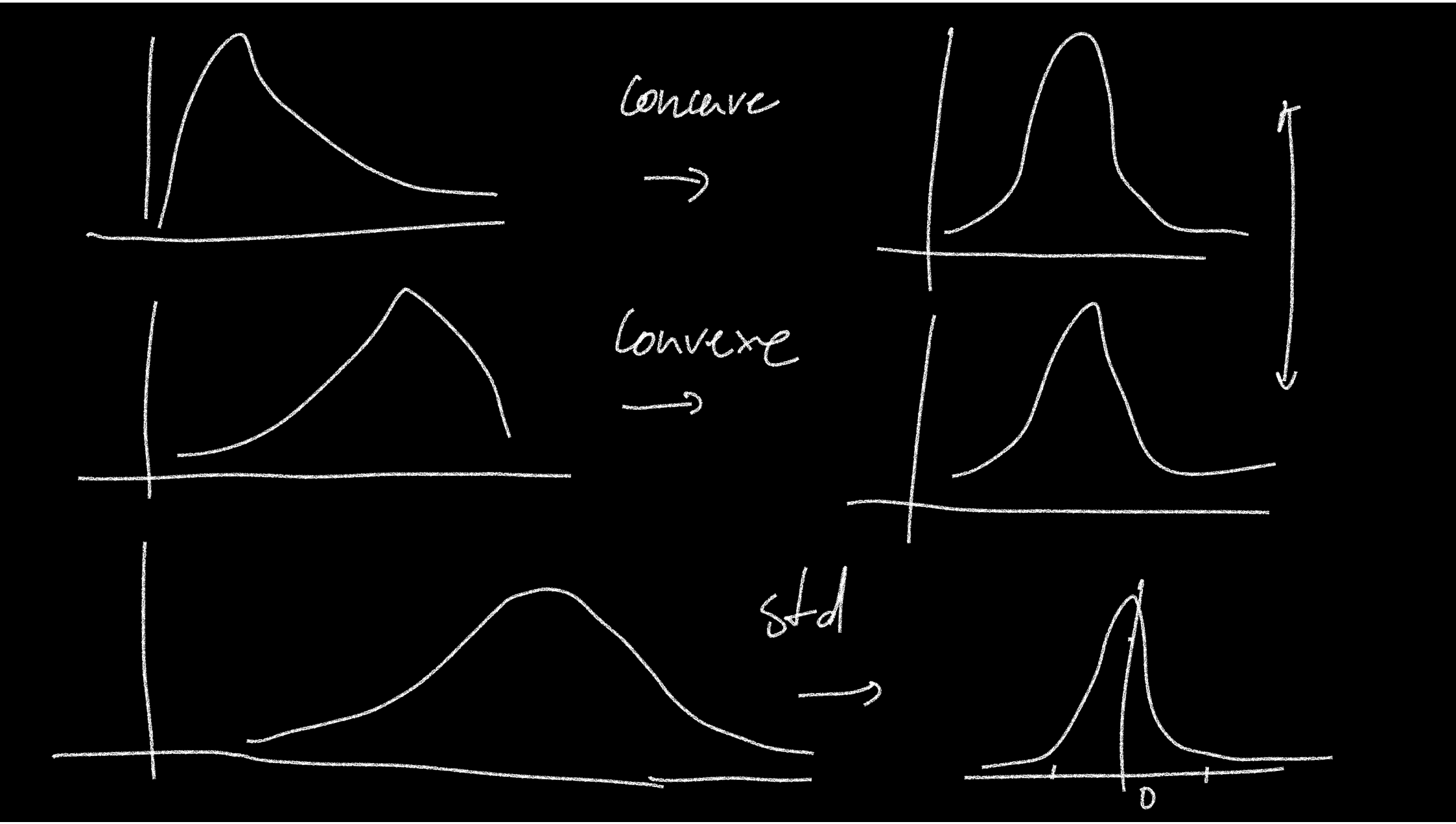
l	L
$\frac{30000-42500}{39157}$	$\frac{42-34}{14}$
$\frac{10000-42500}{39157}$	$\frac{53-34}{14}$
$\frac{100000-42500}{39157}$	$\frac{17-34}{14}$
$\frac{57000-42500}{39157}$	$\frac{24-34}{14}$

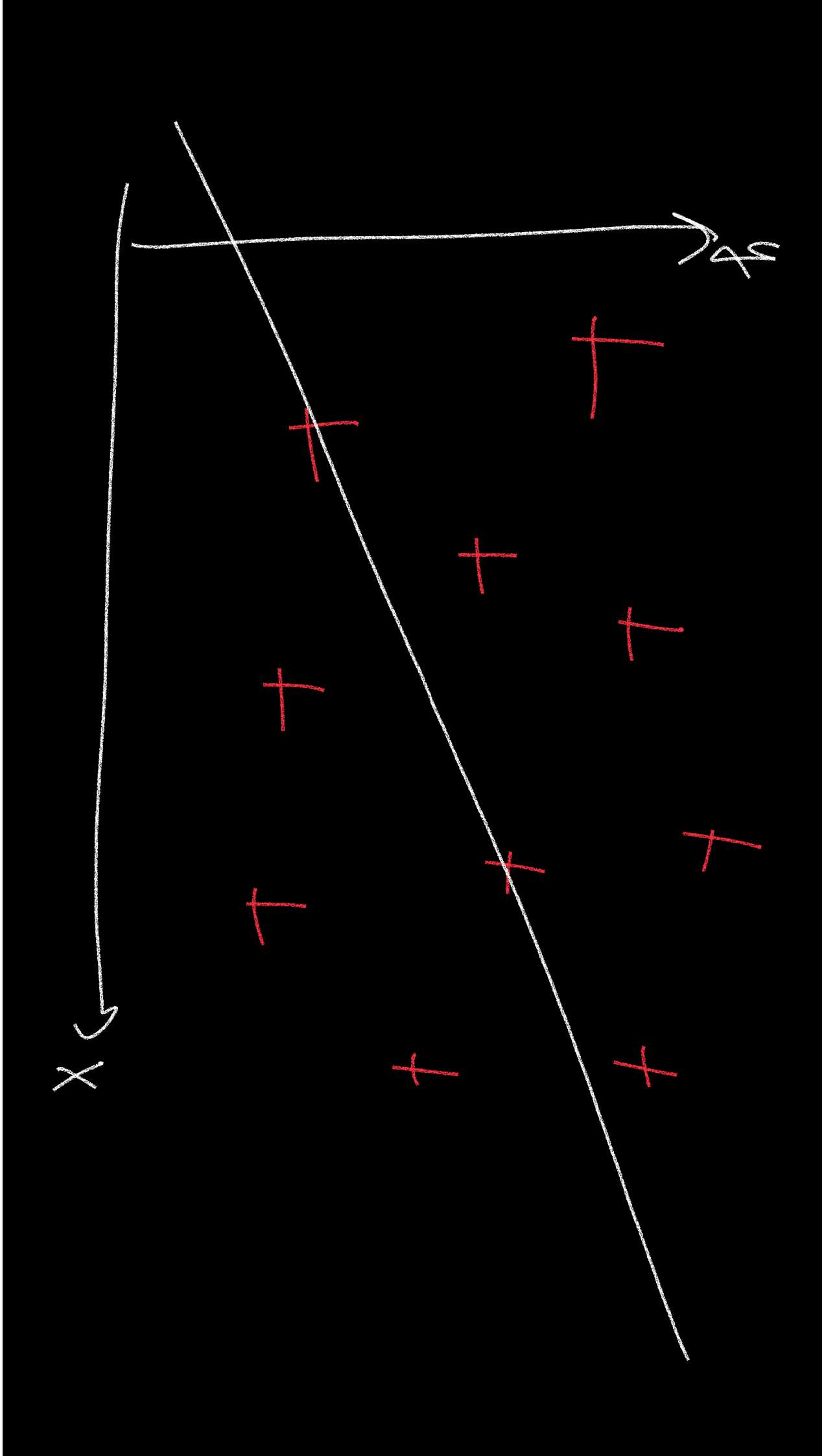
$\underline{l}$

l	L
$-0,33$	$0,8$
-1	1
$1,2$	$-1,2$
$0,33$	$-1,7$

μ	0	0
0	1	1


prix = 3,7 . l + 2,4 . L
effet l > effet L
régression linéaire appliquée
aux colonnes standardisées


Concave
Convexe
std
D


x

Mean Square Error $= \frac{1}{n} \sum \text{erreurs}^2$

↳ $[MSE] = €^2$

Root Mean Square Error $= \sqrt{MSE}$

↳ $[RMSE] = €$

$RMSE = 1\,500\,000\ €$

$RMSE \in [0, +\infty[$

$RMSE \downarrow \Rightarrow$ mieux

Coefficient de détermination $= R^2$

$$= 1 - \frac{\sum_{i=1}^{N} \text{erreur}^2}{\sum_{i=1}^{N} (y[i] - \bar{y}[i])^2}$$

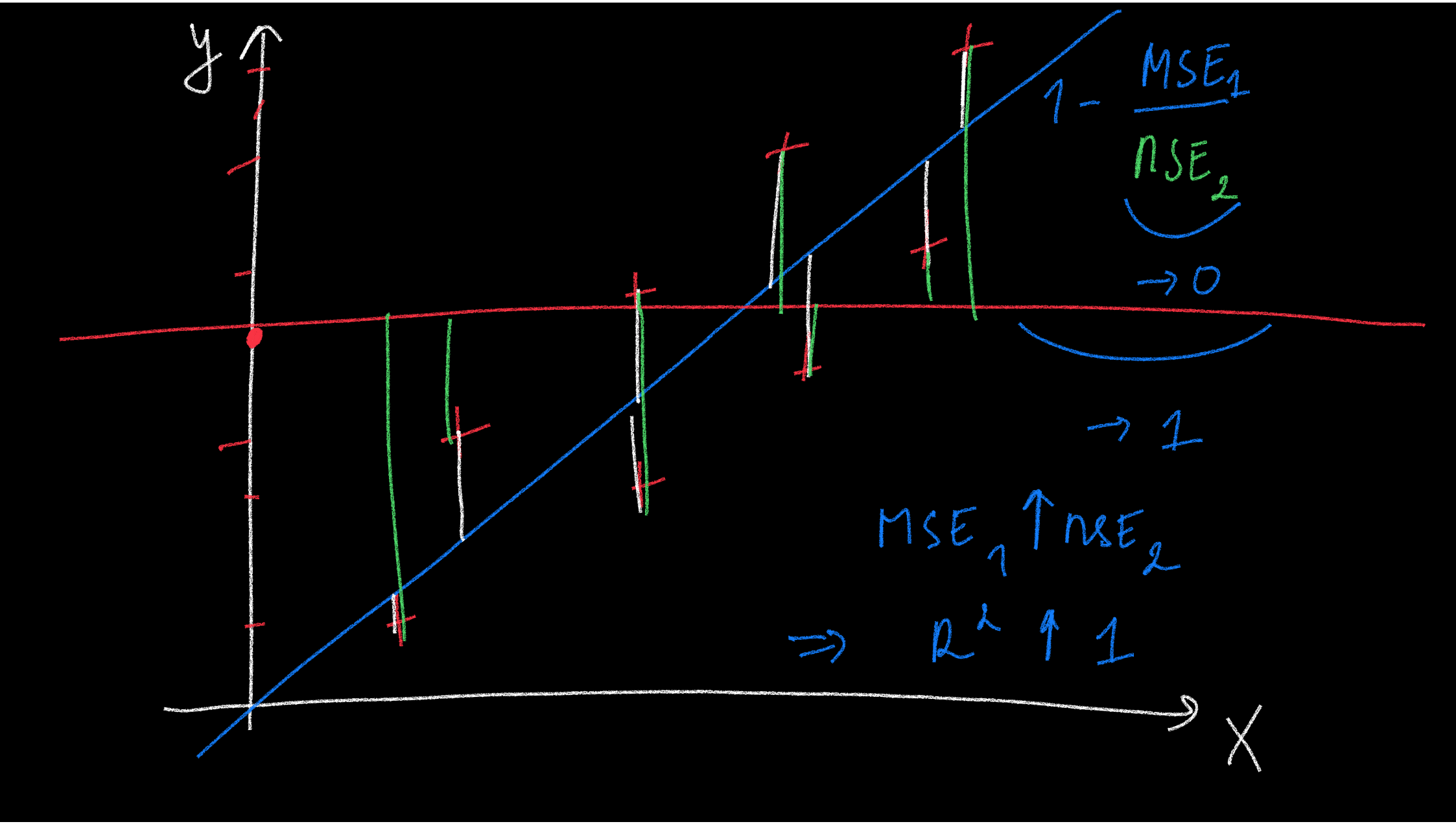
Variance de la R.L. de X

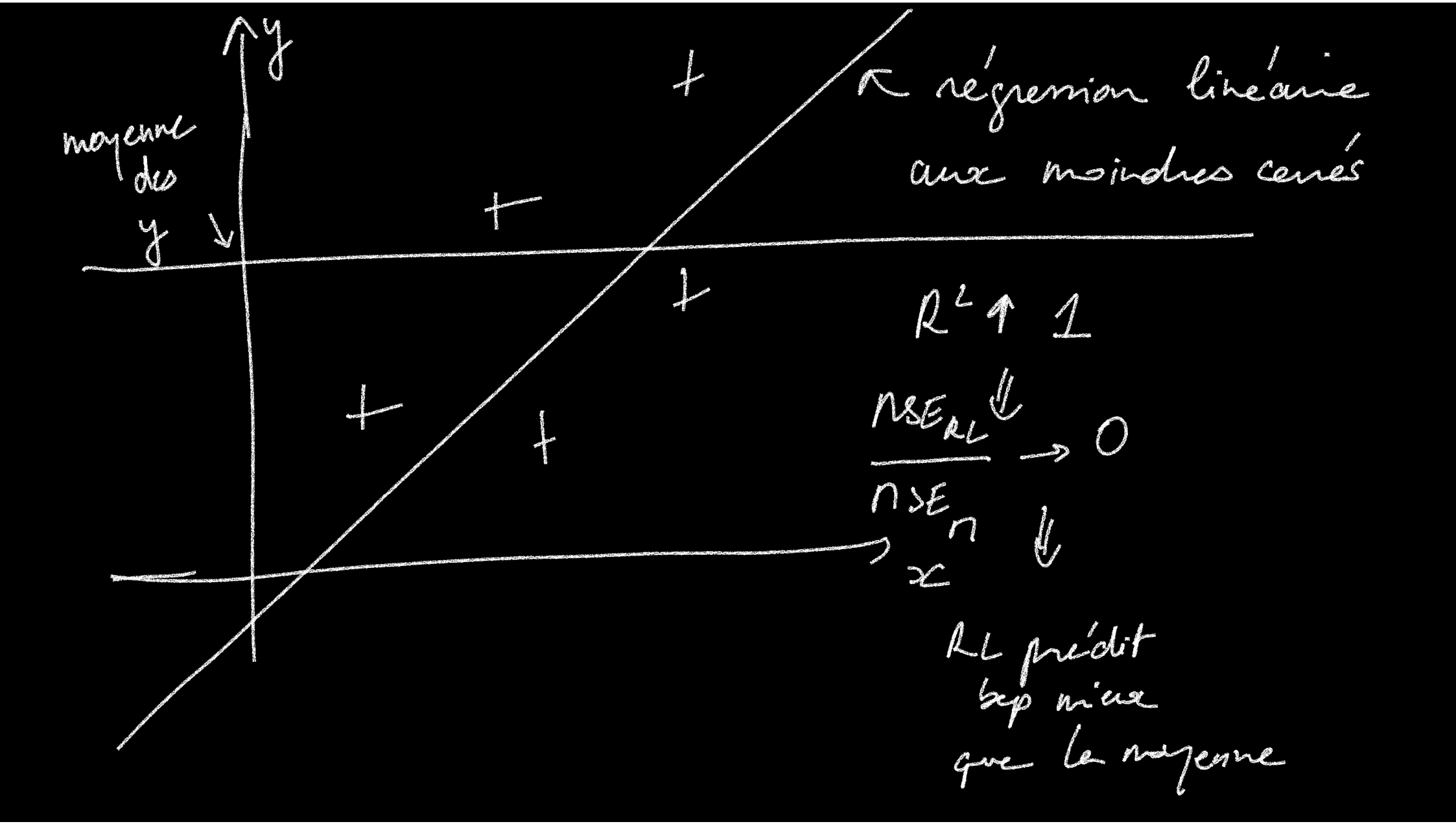
Variance de y

$R^2 \in]-\infty, 1]$

$[0, 1]$

$R^2 \uparrow \Rightarrow$ meilleure régression linéaire


y
X
$1 - \frac{MSE_1}{MSE_2}$
$\to 0$
$\to 1$
$MSE_1 \uparrow MSE_2$
$\Rightarrow R^2 \uparrow 1$

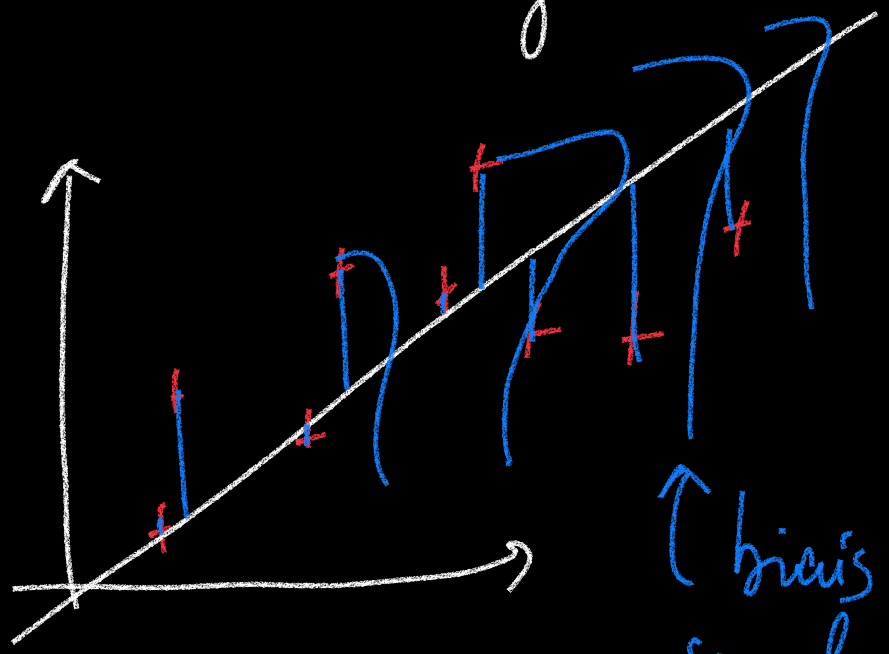

y
moyenne des y
x
régression linéaire aux moindres carrés
$R^2 \uparrow 1$
$\frac{MSE_{RL}}{MSE_m} \to 0$
RL prédit bcp mieux que la moyenne

Coefficient de détermination $= (\text{Coef Corr Pearson})^2$

"Le taux de variance de la variable cible expliqué par la variable explicative."

① ε $\quad y = ax + b + \varepsilon$

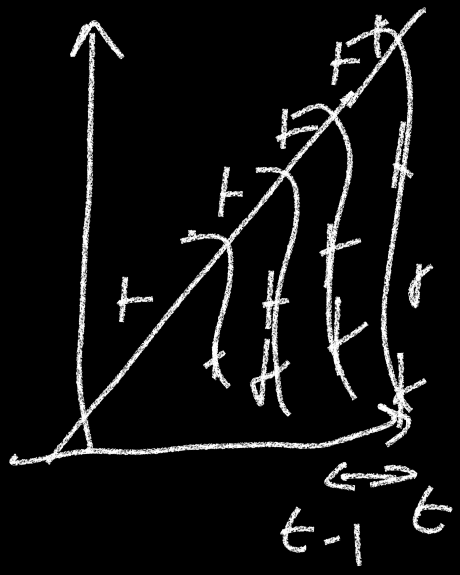
$\varepsilon \sim$ gaussien



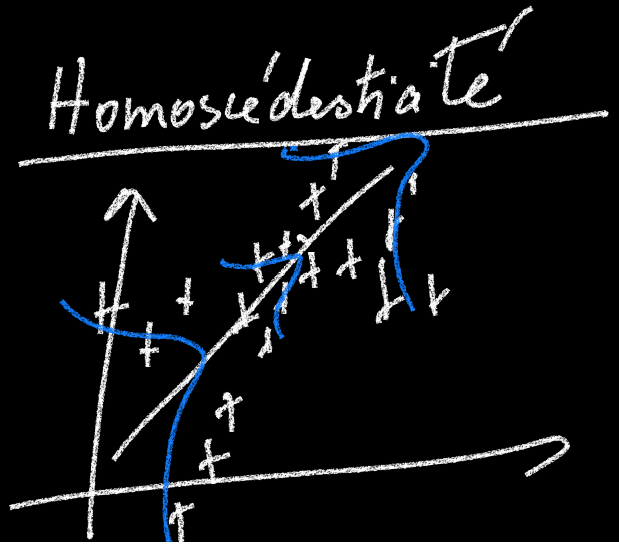
↑ biais sur la droite

⇝ bruit non gaussien

② bruit indépendant



③ Homoscédasticité



$y = x_1, \dots, x_p$

Variables colinéaires

$T°C$, $T°K$, $T°F$

On veut des variables explicatives les Θ redondantes possibles.