

---

# **Partie 1 Régression simple**

## **Modèle ?**

Représentation des phénomènes  
en réalité en vue de comprendre  
le fonctionnement,

---

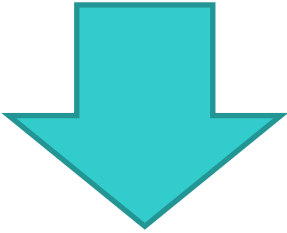
Marketing direct en ligne:

construire un modèle pour identifier les clients les plus susceptibles d'acheter des produits de leur prochain catalogue

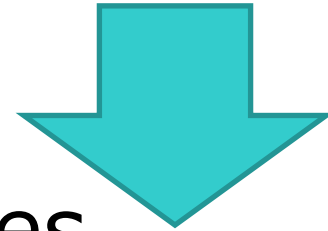
Clients identifiés par le modèle comme ayant peu de chance d'acheter seront exclu de la prochaine liste d'envoi.

---

# Modèles mathématiques et statistiques



Modèles  
déterministes



Modèles  
probabilistes

---

# Modèles mathématiques et statistiques

- ❑ Déterministes
- ❑ Random Error  
(aléatoires)

---

# Modèles mathématiques et statistiques

Exemple:

Ventes d'un produit = frais de  
publicité + force de vente + prix +  
,,,+ erreur de perturbation

---

**Probabilistic  
Models**

**Regression  
Models**

**Correlation  
Models**

**Other  
Models**

```
graph TD; A[Probabilistic Models] --> B[Regression Models]; A --> C[Correlation Models]; A --> D[Other Models];
```

# Corrélation et régression linéaire simple

---

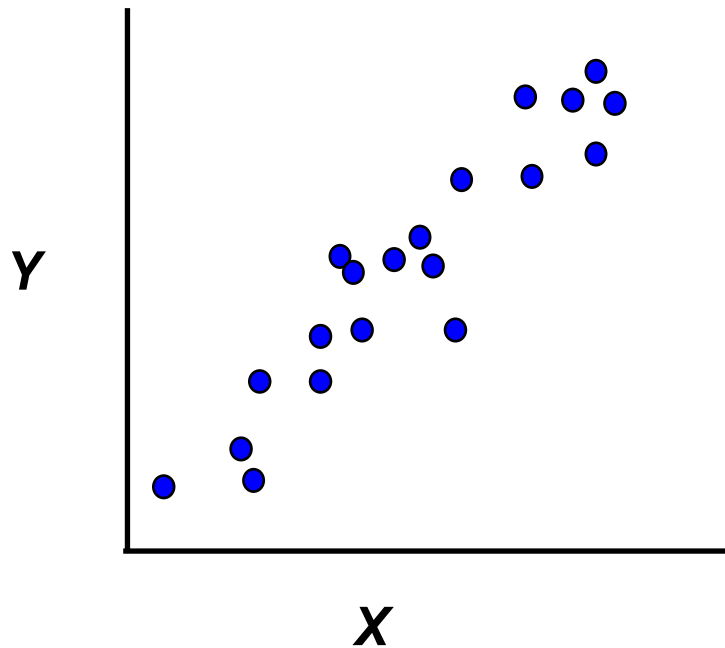
- ❑ La corrélation
- ❑ La régression linéaire simple

# Introduction

---

## Etude de la relation entre deux variables quantitatives:

Nuage de points:



- description de l'association linéaire: corrélation, régression linéaire simple

- explication / prédiction d'une variable à partir de l'autre: modèle linéaire simple



# La corrélation

---

Statistique descriptive de la relation entre X et Y: variation conjointe

## 1. La covariance

Dans l'échantillon:

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}$$

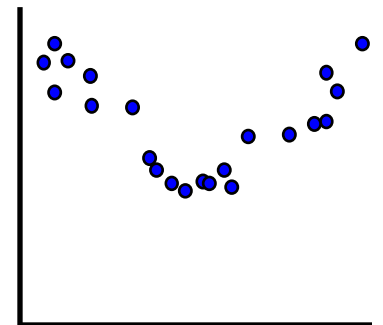
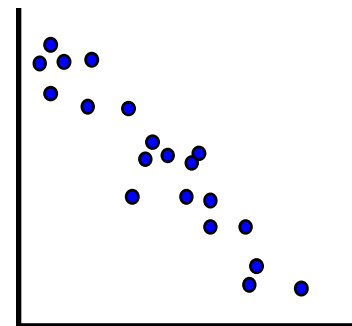
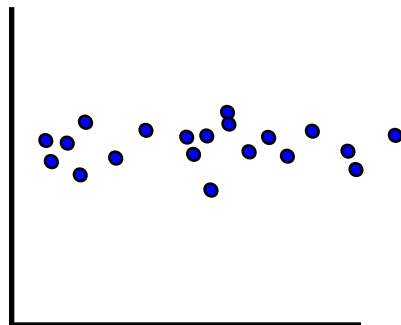
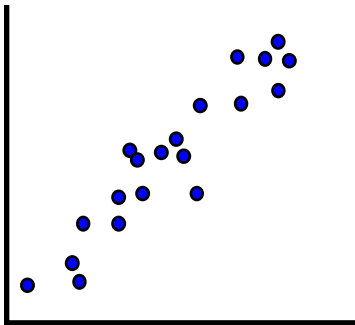
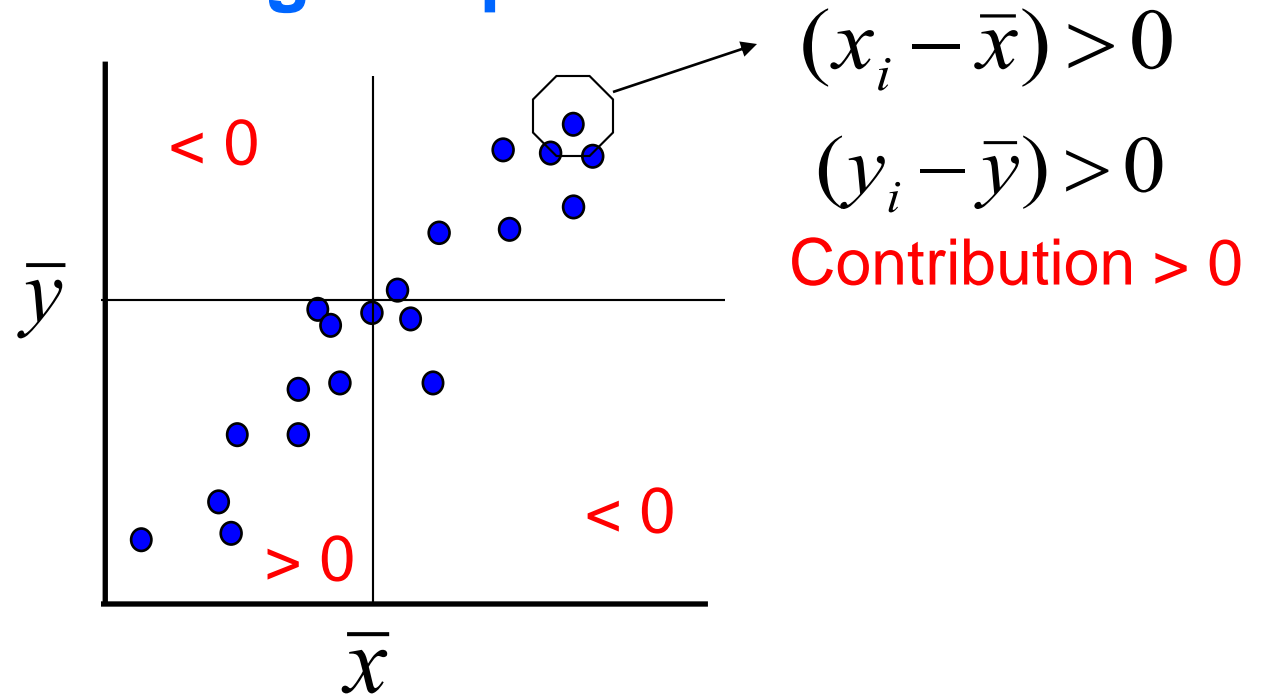
Estimation pour la population:

$$\text{cov}(x, y) = \hat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n x_i y_i - \frac{n}{n-1} \bar{x} \bar{y}$$

# La corrélation

## Covariance et nuage de points



# La corrélation

---

## 2. Le coefficient de corrélation linéaire

« de Pearson »

$$\frac{\text{Cov}(x,y)}{(\text{écart type de } x * \text{écart type de } y)} = \frac{\text{cov}(x,y)}{\delta x \delta y}$$

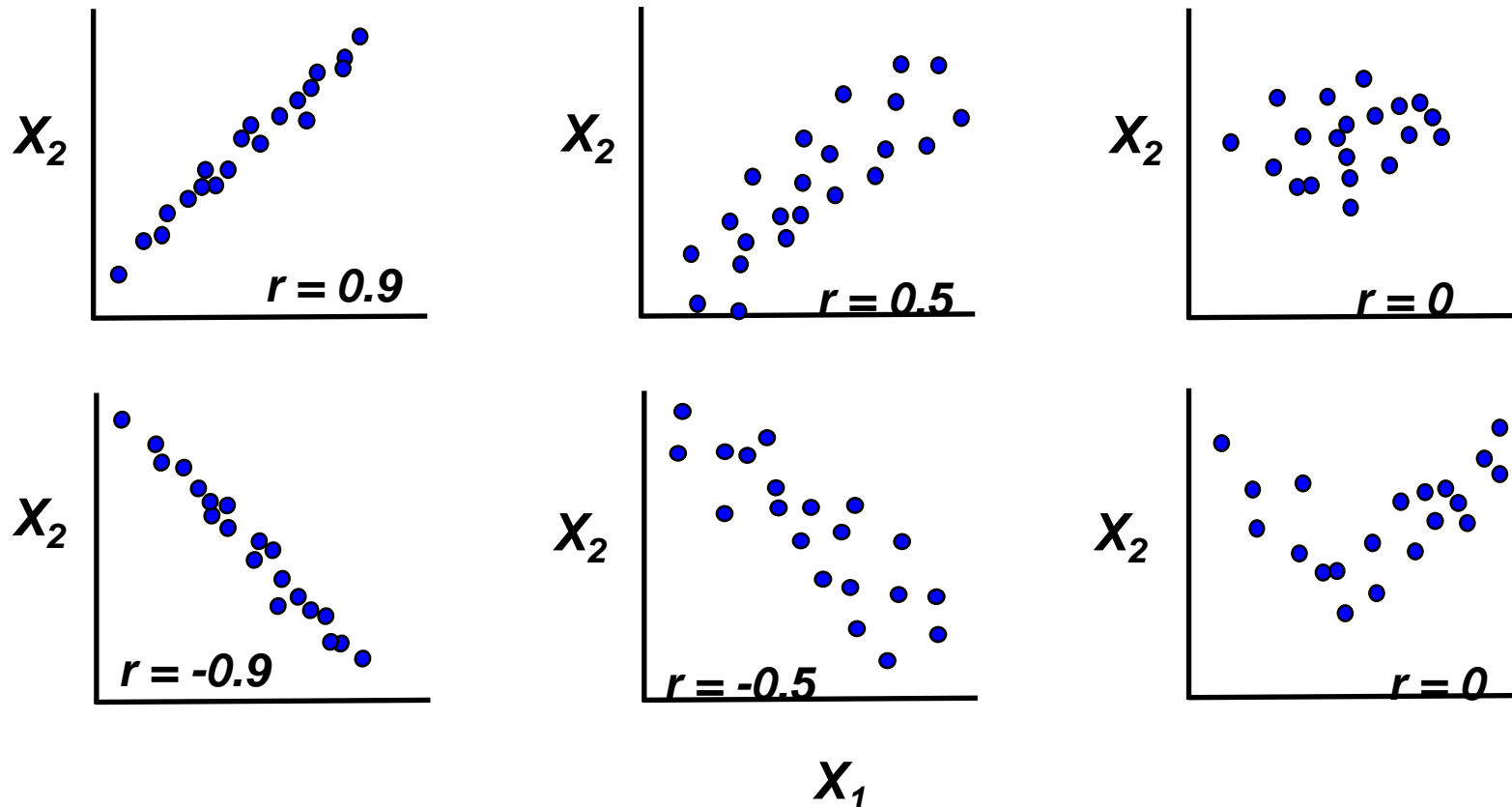
$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}}$$

# La corrélation

---

## 2. Le coefficient de corrélation linéaire

Indice de covariance absolu:  $-1 \leq r \leq 1$



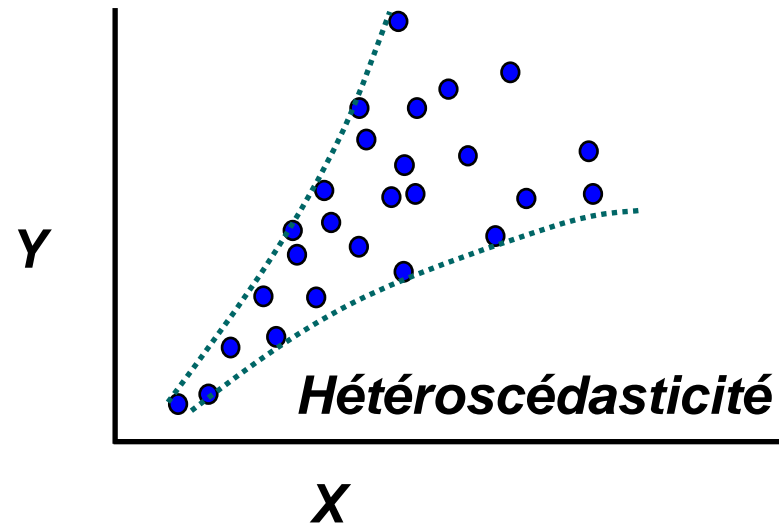
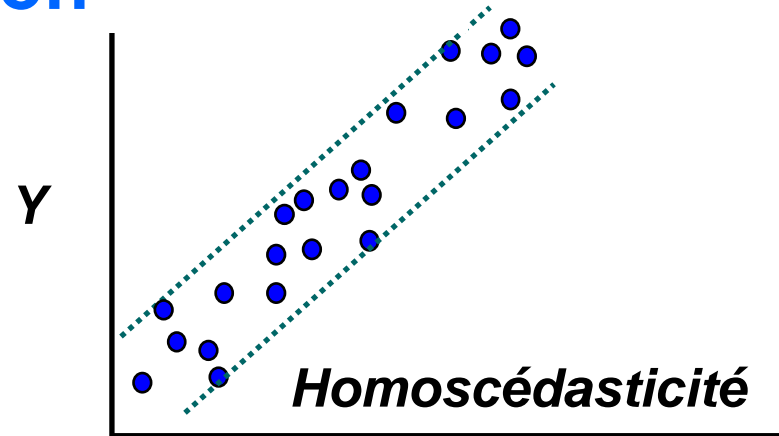
# La corrélation

---

## 3. Conditions d'utilisation

### Homoscédasticité

La variance de  $Y$  est indépendante de  $X$  et vice-versa.



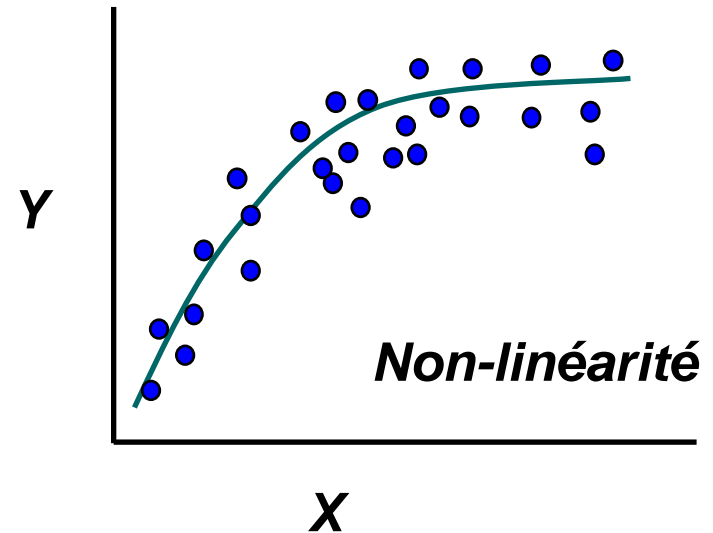
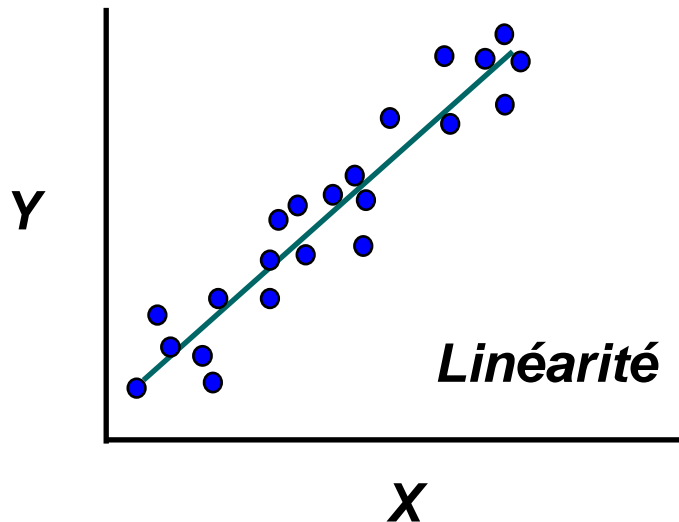
# La corrélation

---

## 3. Conditions d'utilisation

### Linéarité

La relation est linéaire



**1 Explanatory  
Variable**

## **Regression Models**

**2+ Explanatory  
Variables**

**Simple**

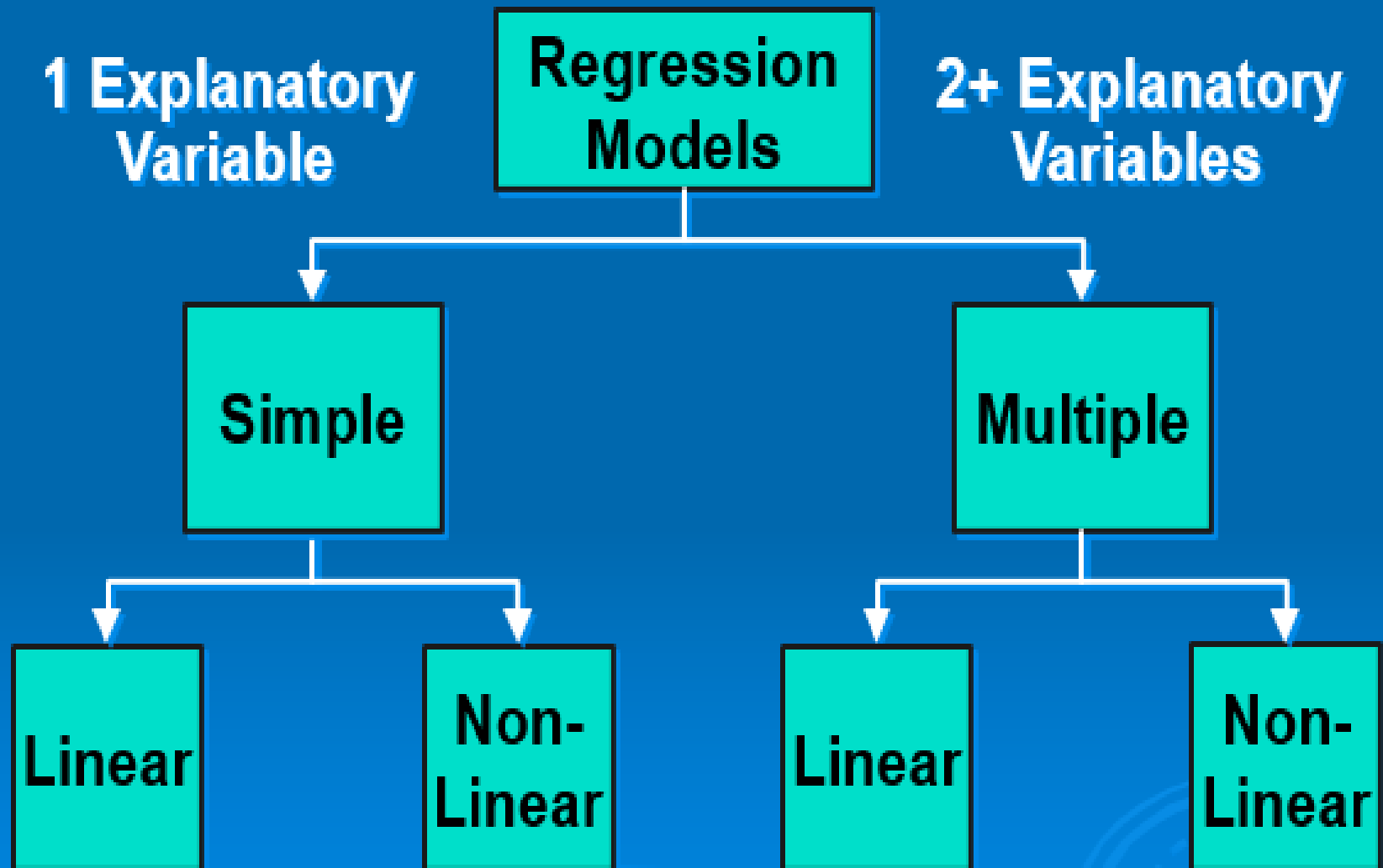
**Multiple**

**Linear**

**Non-  
Linear**

**Linear**

**Non-  
Linear**



# La régression linéaire simple

---

## 1. Le modèle

On suppose:  $y = f(x) = a + bx$

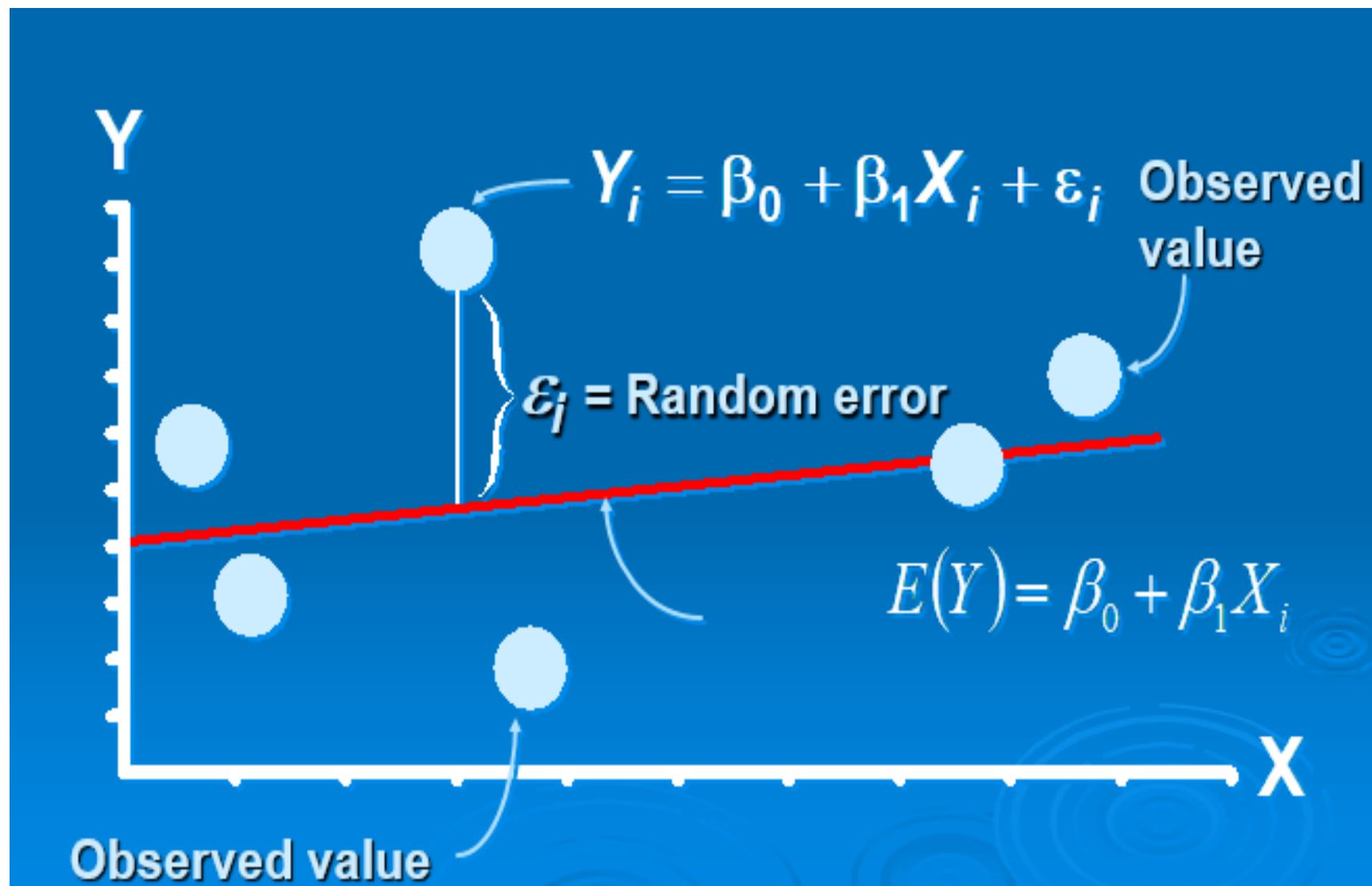
Modèle:  $Y_i = a + bX_i + e_i$

avec, pour  $X = x_i$ ,  $Y_i : N(a+bx_i, \sigma)$

$X$  = variable explicative  
(« indépendante »), **contrôlée**  
 $Y$  = variable expliquée  
(dépendante ), **aléatoire**





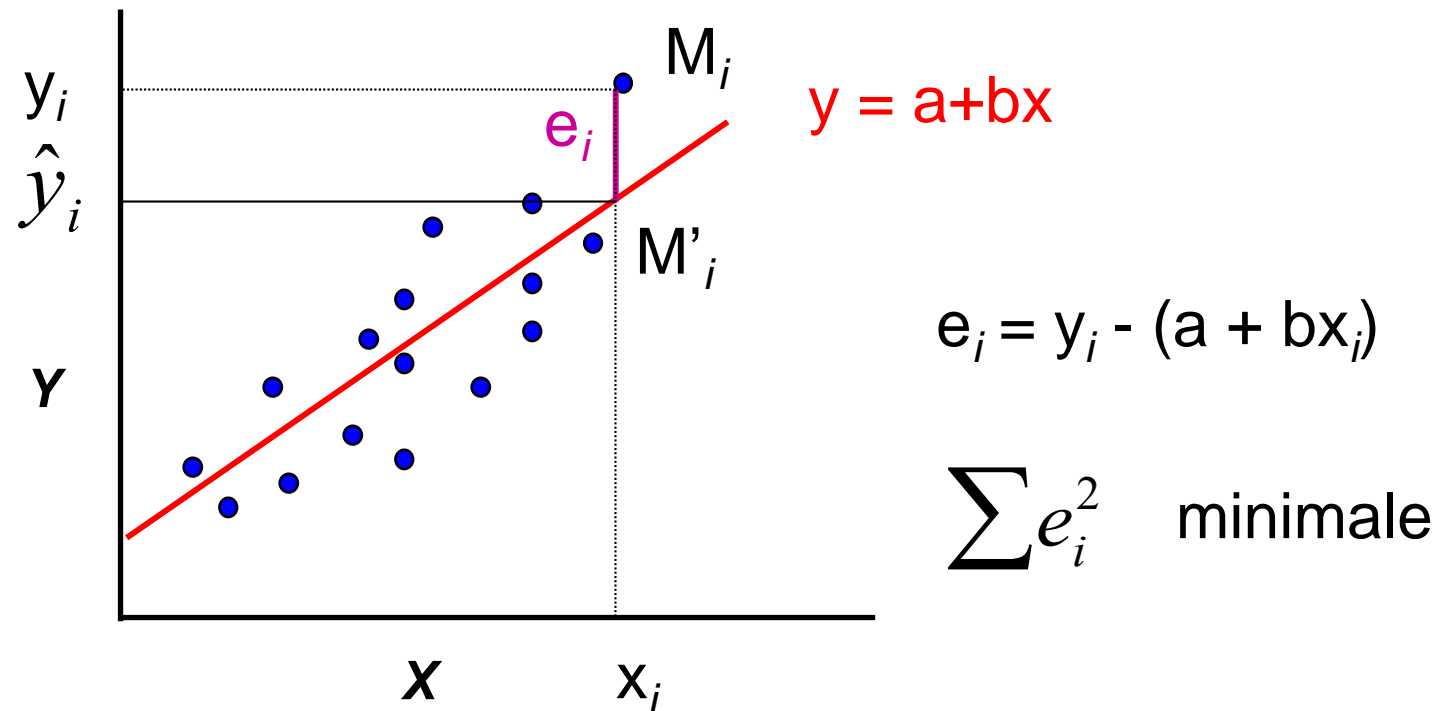


# La régression linéaire simple

## 2. L'estimation des paramètres

a? b?

Méthode d'estimation: les moindres carrés:



# La régression linéaire simple

---

## 2. L'estimation des paramètres

### Méthode des moindres carrés

On cherche le minimum de  $\sum_{i=1}^n (y_i - (a + bx_i))^2 = E(a, b)$

$$\left\{ \begin{array}{l} \frac{\partial E}{\partial a} = \sum_{i=1}^n 2(y_i - (a + bx_i))(-1) = 0 \quad (1) \\ \frac{\partial E}{\partial b} = \sum_{i=1}^n 2(y_i - (a + bx_i))(-x_i) = 0 \quad (2) \end{array} \right.$$

# La régression linéaire simple

---

## 2. L'estimation des paramètres

### Méthode des moindres carrés

$$(1) \Rightarrow \sum_{i=1}^n y_i = \sum_{i=1}^n (a + bx_i) = na + b \sum_{i=1}^n x_i$$

$$n\bar{y} = na + nb\bar{x}$$

$$a = \bar{y} - b\bar{x}$$

# La régression linéaire simple

---

## 2. L'estimation des paramètres

### Méthode des moindres carrés

$$n(\text{cov}(x, y) + \bar{x}\bar{y}) - (\bar{y} - b\bar{x})n\bar{x} - bn(s_x^2 + \bar{x}^2) = 0$$

$$\text{cov}(x, y) = bs_x^2 \quad b = \frac{\text{cov}(x, y)}{s_x^2}$$

$$\text{Si } y = a + bx \text{ alors } \hat{b} = \frac{\text{cov}(x, y)}{s_x^2} \quad \text{et} \quad \hat{a} = \bar{y} - b\bar{x}$$

On peut alors prédire  $y$  pour  $x$  compris dans l'intervalle des valeurs de l'échantillon:  $\hat{y}_i = \hat{a} + \hat{b}x_i$

# La régression linéaire simple

---

## 3. Qualité de l'ajustement

On a supposé:  $Y_i = a + bX_i + e_i$  avec

pour  $X = x_i$ ,  $Y_i : N(a+bx_i, \sigma)$

- distribution normale des erreurs
- variance identique (homoscédasticité)
- indépendance:  $\text{cov}(e_i, e_j) = 0$
- linéarité de la relation

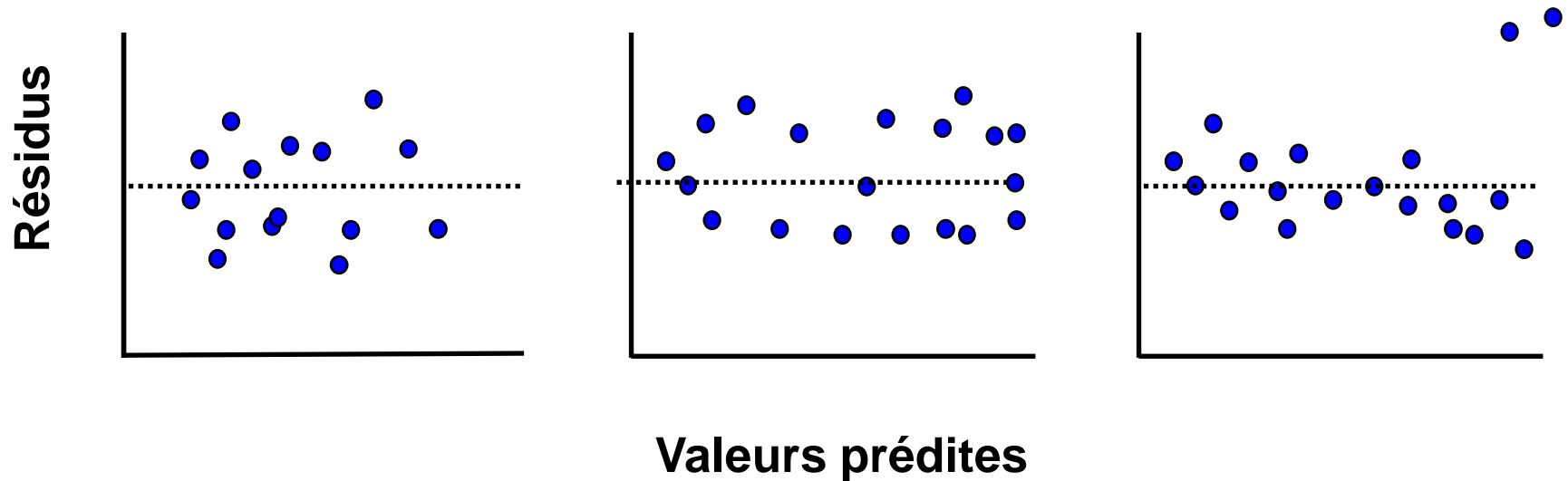
Test *a posteriori* : étude du nuage de points/ du graphe des résidus

# La régression linéaire simple

---

## 3. Qualité de l'ajustement

### Normalité de l'erreur



Questions à se poser: structure de l'erreur?

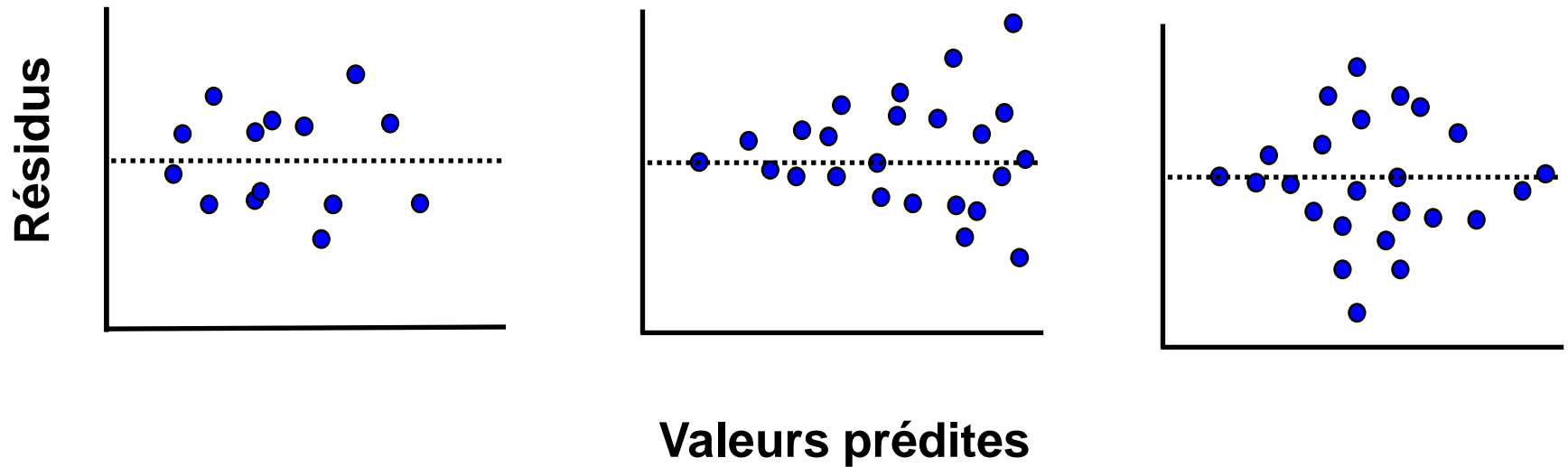
Valeurs extrêmes: ont-elles un sens? Influent-elles l'estimation des paramètres?

# La régression linéaire simple

---

## 3. Qualité de l'ajustement

### Homoscédasticité



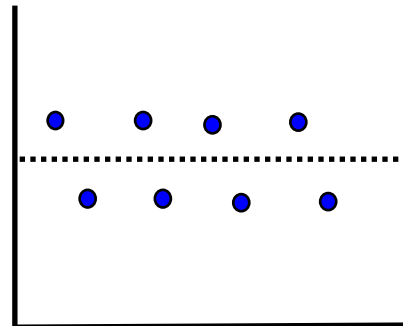
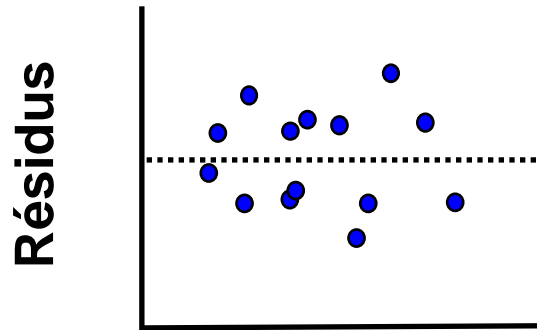


# La régression linéaire simple

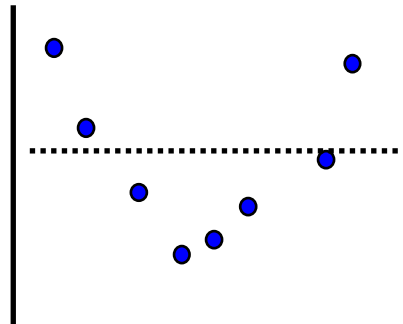
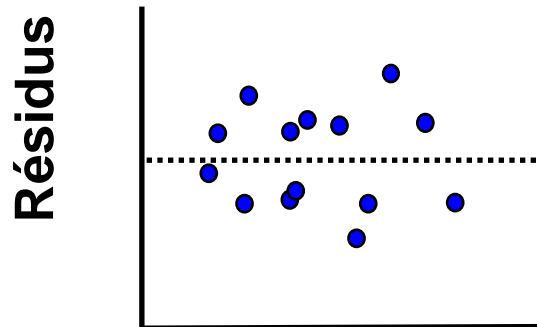
---

## 3. Qualité de l'ajustement

### Indépendance entre erreurs, linéarité



Structure de l'erreur?



Relation non linéaire?

# La régression linéaire simple

---

## 4. Coefficient de détermination

### Décomposition de la variation

Quelle part de la variabilité de Y est expliquée par la relation linéaire avec X?

Variabilité? Somme des Carrés des Ecartes SCE:

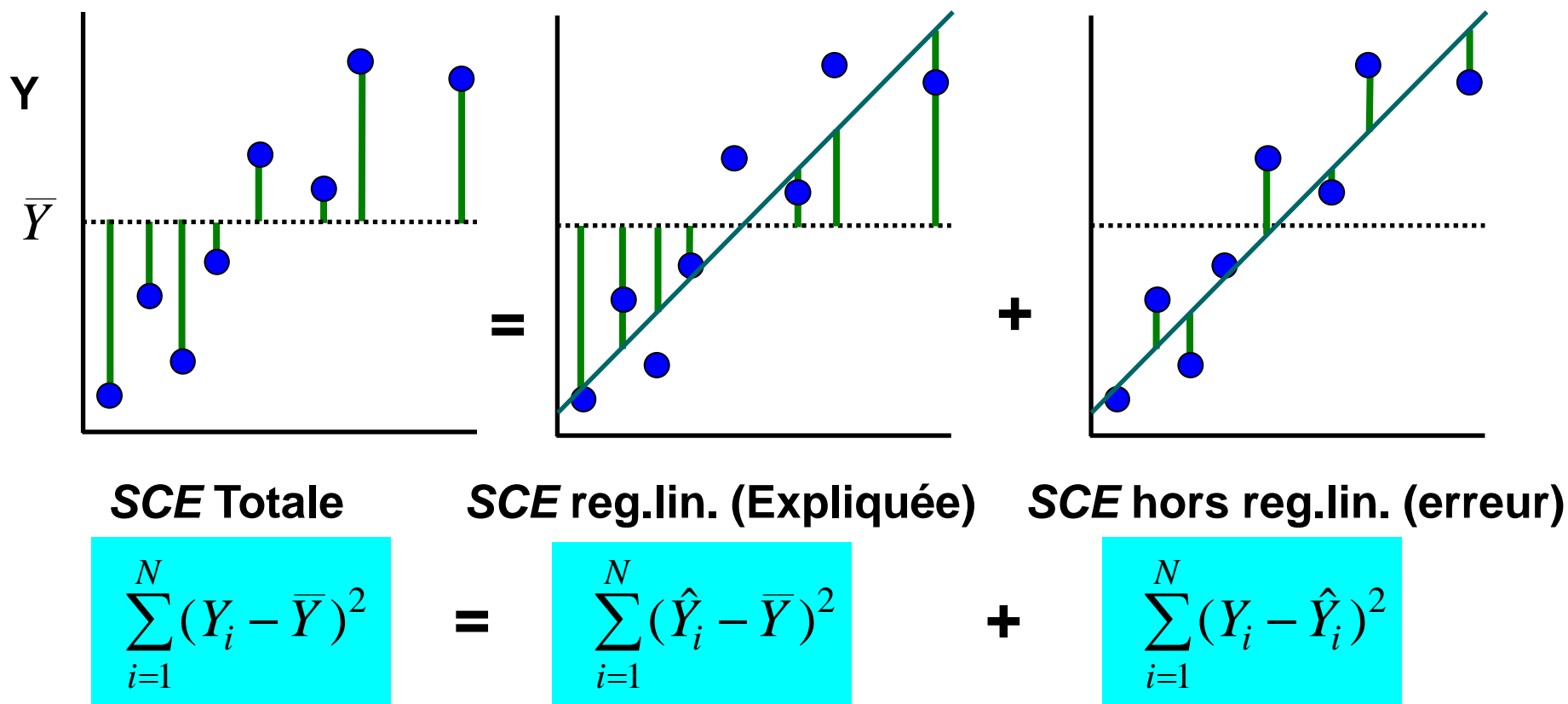
**Variance  
totale**

$$SCE_T = \sum_{i=1}^n (y_i - \bar{y})^2 = ns_y^2$$

# La régression linéaire simple

## 4. Coefficient de détermination

### Décomposition de la variation



# La régression linéaire simple

---

## 4. Coefficient de détermination

La décomposition de la SCE permet d'estimer la part de SCE de Y expliquée par la régression:

Coefficient de détermination

$$r^2 = \frac{SCE_{reg.lin.}}{SCE_T} \quad \frac{\textit{var expliquée}}{\textit{var totale}}$$

$$0 \leq r^2 \leq 1$$

Relation avec r?

# La régression linéaire simple

---

## 4. Coefficient de détermination

### Relation entre $r$ et $r^2$

$$\begin{aligned} SCE_{reg.lin.} &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n ((a + bx_i) - (a + b\bar{x}))^2 \\ &= b^2 \sum_{i=1}^n (x_i - \bar{x})^2 = b^2 ns_x^2 = b^2 SCE_x \end{aligned}$$

$$\text{Donc } r^2 = \frac{b^2 ns_x^2}{ns_y^2} = \left( \frac{\text{cov}(x, y)}{s_x^2} \right)^2 \frac{s_x^2}{s_y^2} = \frac{(\text{cov}(x, y))^2}{s_x^2 s_y^2} = (r)^2$$

En particulier,  $r = 0 \iff r^2 = 0$

# La régression linéaire simple

---

## 5. Tests

Test de la décomposition de la variation ou analyse de variance (ANOVA):  $H_0 : a=b = 0$

$$\frac{\sigma_{reg.lin.}^2}{\sigma_{horsreg.lin.}^2} = \frac{SCE_{reg.lin.} / 1}{SCE_{horsreg.lin.} / (n - 2)} : F_{n-2}^1$$

Si  $F(\text{obs}) > F_c$  (tabulée refuser  $H_0$ )  
Si non accepter  $H_0$

---

Test de significativité par variable:

$$\frac{\hat{a} - a}{\hat{\sigma}_{\hat{a}}} \equiv \mathfrak{T}(n - 2)$$

$$\frac{\hat{b} - b}{\hat{\sigma}_{\hat{b}}} \equiv \mathfrak{T}(n - 2)$$

Ho: Coef a = 0

H1: coef a  $\neq$  0

Si  $T(\text{obs}) = \text{coef}/\text{écart type estimé}$   
> tc (tabulée) refuser H0  
Si non accepter Ho

---

Ventes en fonction des frais de publicité:

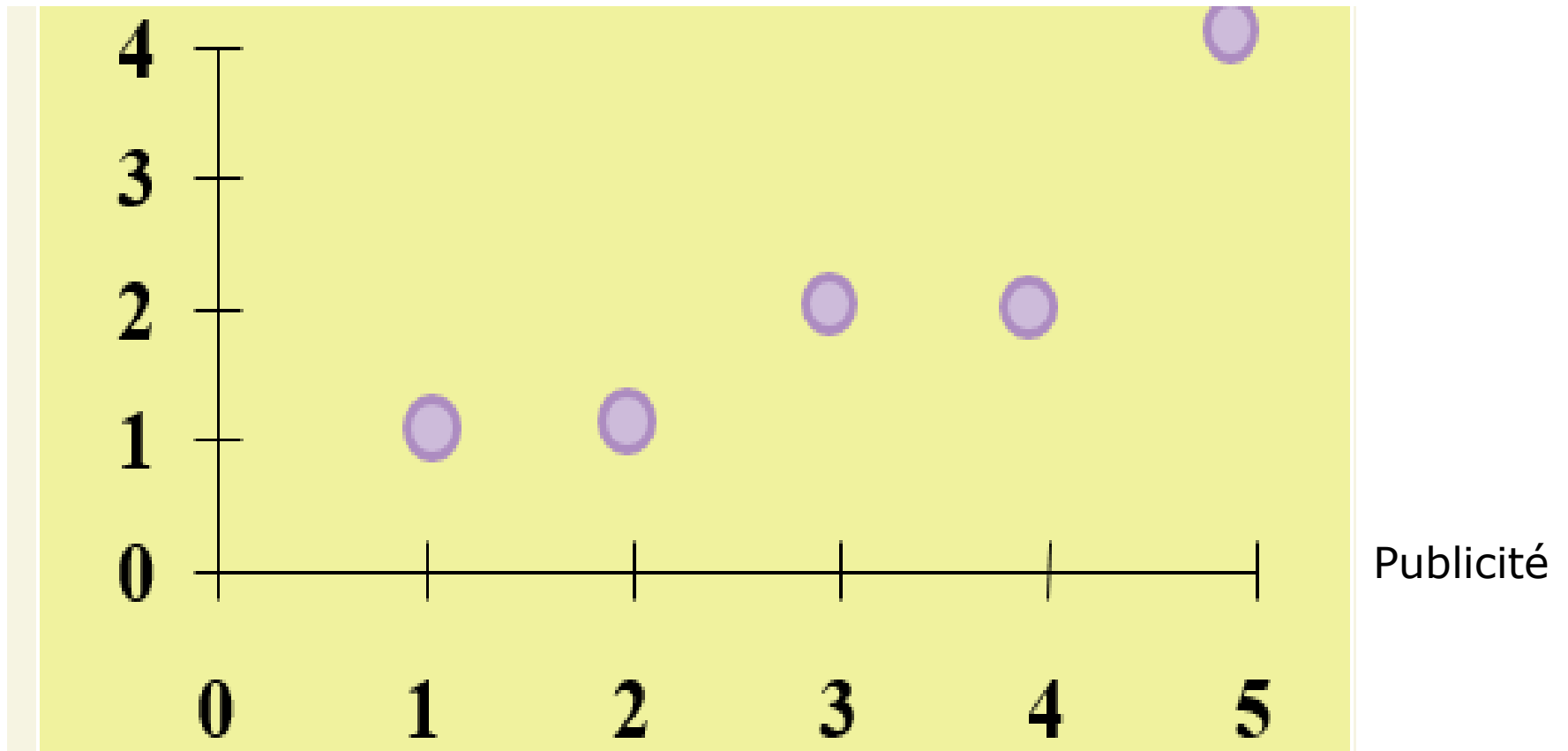
<u>pub</u>	<u>ventes (Units)</u>
1	1
2	1
3	2
4	2
5	4





---

Ventes



---

$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

# Paramètres Estimés

---

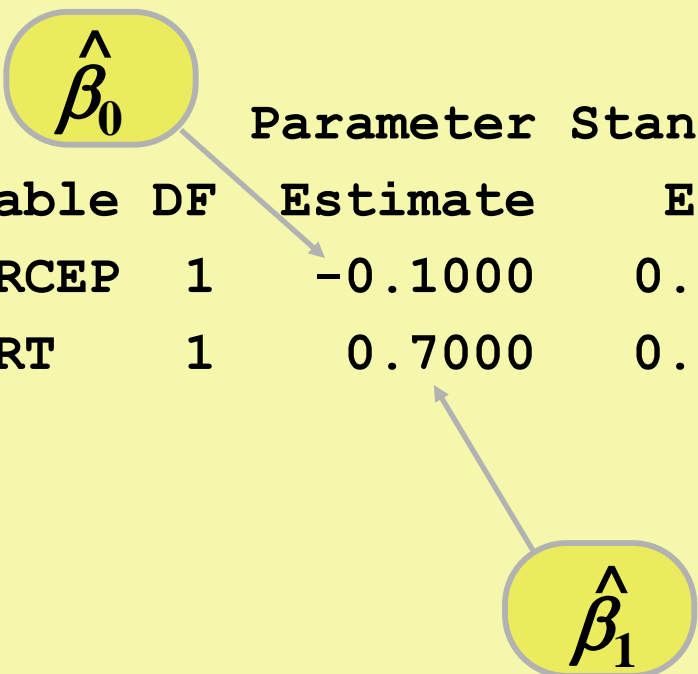
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = .70$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2 - (.70)(3) = -.10$$

$$\hat{y} = -.1 + .7x$$

# Résultats

## Parameter Estimates



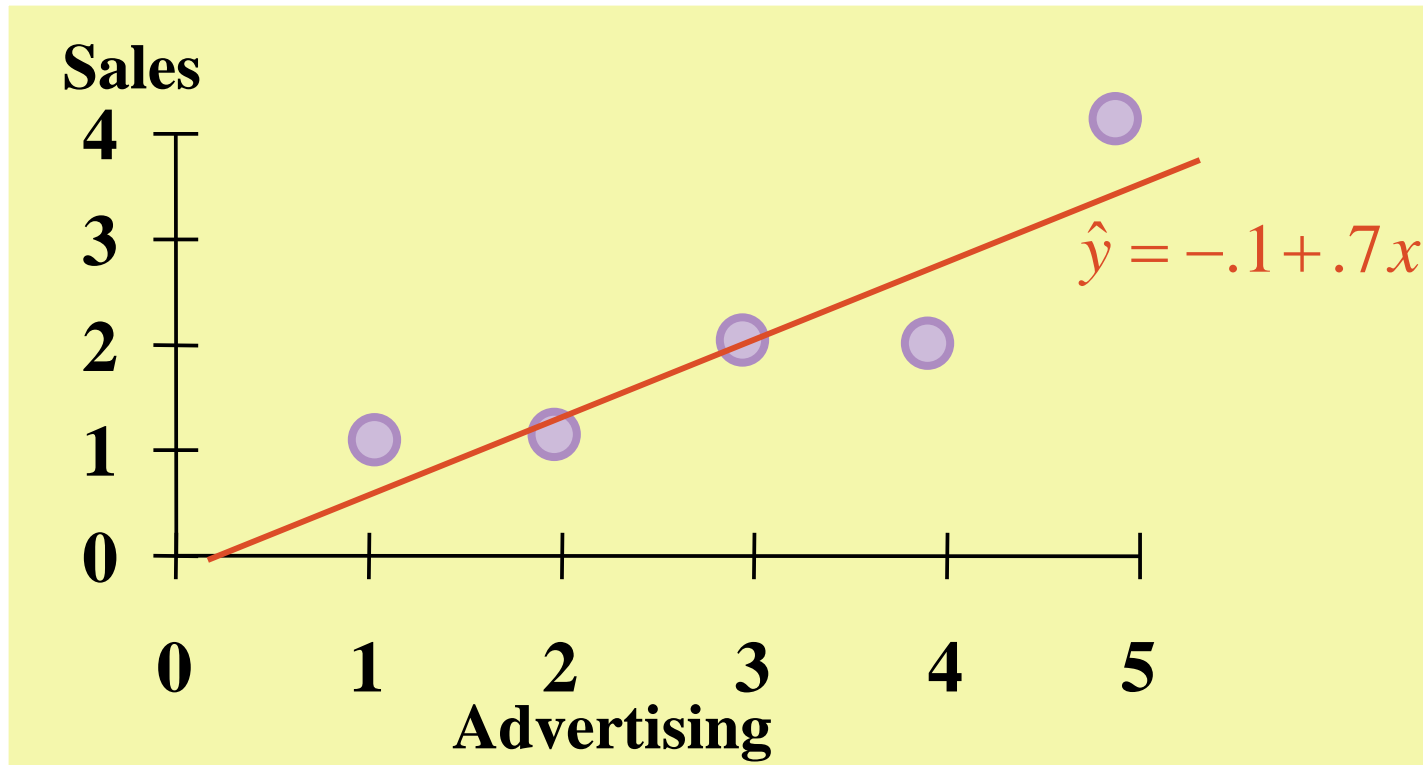
Parameter Estimates					
Parameter Standard T for H0:					
Variable	DF	Estimate	Error	Param=0	Prob> T
INTERCEP	1	-0.1000	0.6350	-0.157	0.8849
ADVERT	1	0.7000	0.1914	3.656	0.0354

$$\hat{y} = -.1 + .7x$$

# Regression Line Fitted

---

## (representation de Y estimée ou ajustée)



---

## Exercise

Y	X
16	20
18	24
23	28
24	22
28	32
29	28
26	32
31	36
32	41
34	41

## 2. Analyse de regression – relation exponentielle

---

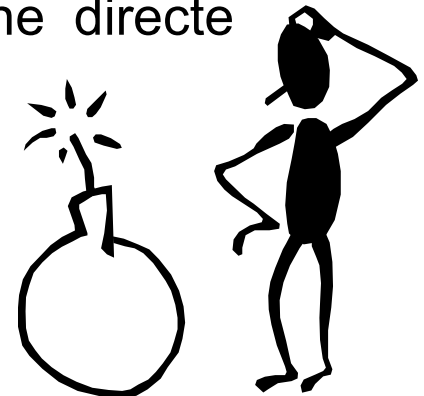
La fonction exponentielle est très courante en sciences

$$y = ae^{bx}$$

Par exemple la décroissance d'un bien ...

Si les constantes  $a$  et  $b$  sont inconnues, on espère pouvoir les estimer à partir de  $x$  et  $y$ . Malheureusement l'approche directe fournit des équations insolubles.

Alors... comment faire????



## 2. Analyse de regression – relation exponentielle

---

Très facile! On transforme l'équation non linéaire en une équation linéaire. Linéarisation en prenant le logarithme:

$$\ln y = \ln a + bx$$

$\ln y$  devient linéaire en  $x$



## 2. Analyse de regression – relation exponentielle

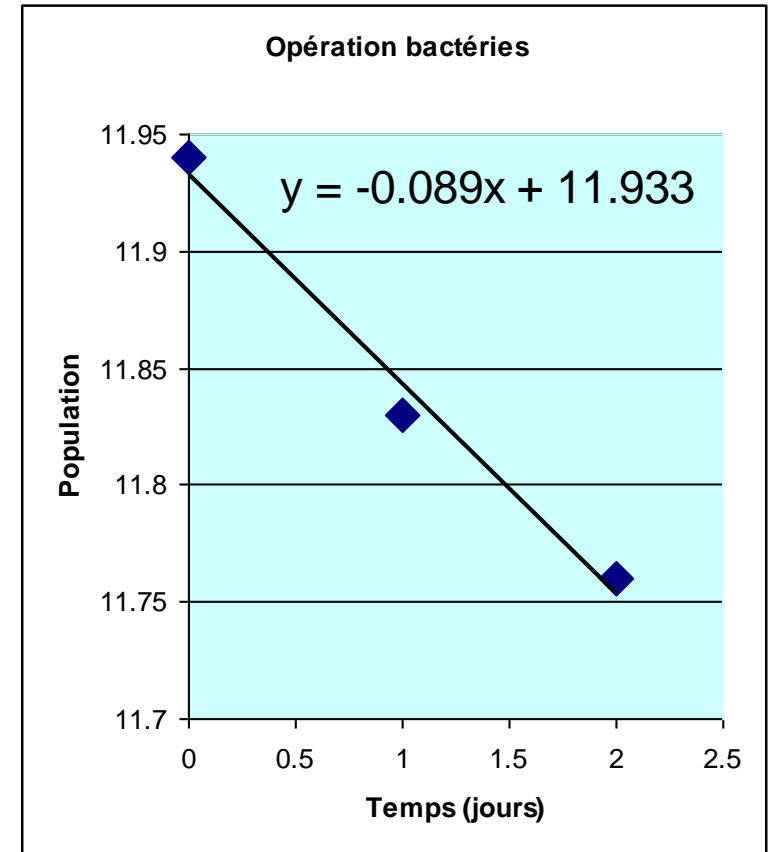
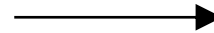
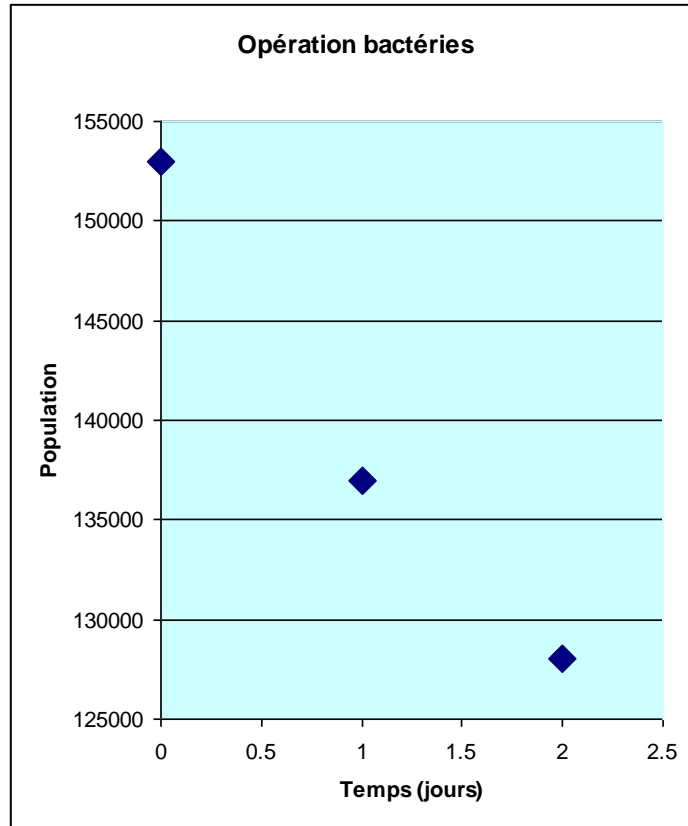
Une population de bactéries décroît exponentiellement:

$$N = N_0 e^{-t/\tau}$$

$t$  est le temps et  $\tau$  est la vie moyenne de la population. A rapprocher de la demi-vie  $t_{1/2}$ ; en fait  $t_{1/2} = (\ln 2) \tau$ .

Temps $t_i$ (jours)	Population $N_i$	$Z_i = \ln N_i$
0	153000	11.94
1	137000	11.83
2	128000	11.76

## 2. Analyse de regression – relation exponentielle



$$\ln N_0 = 11,93 \text{ et } (-1/\tau) = -0.089 \text{ j}^{-1}$$
$$\tau = 11,2 \text{ jours}$$

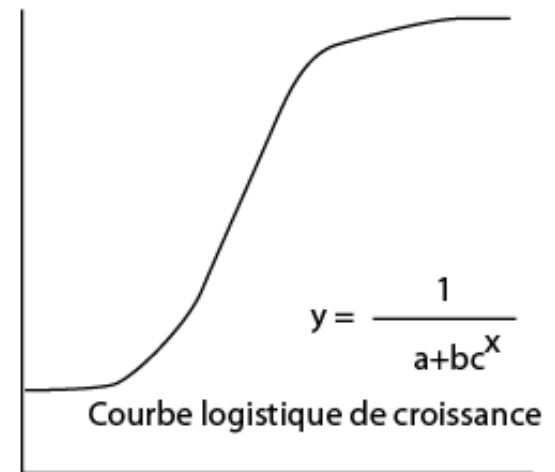
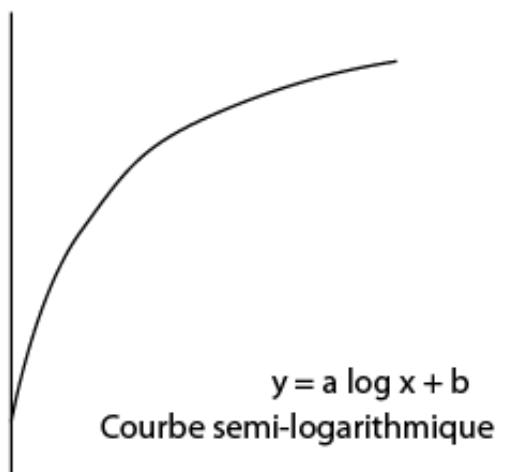
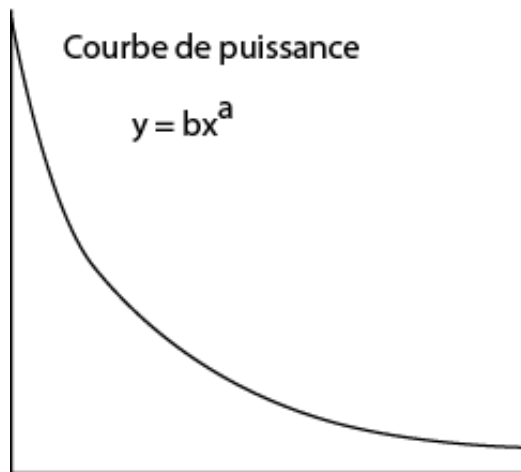
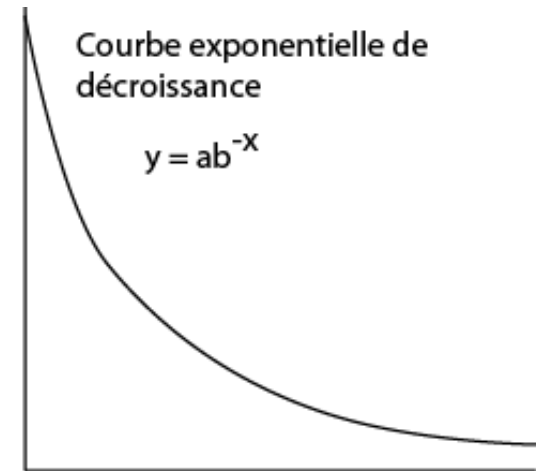
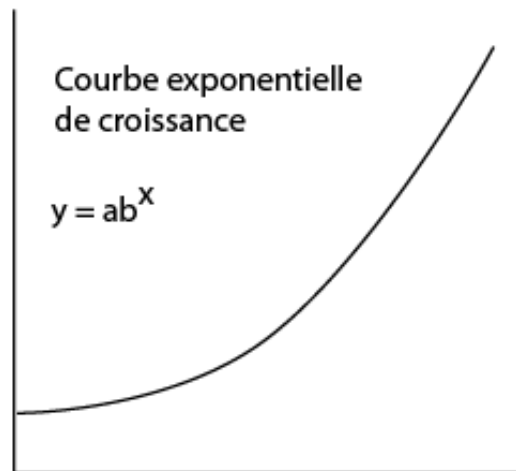
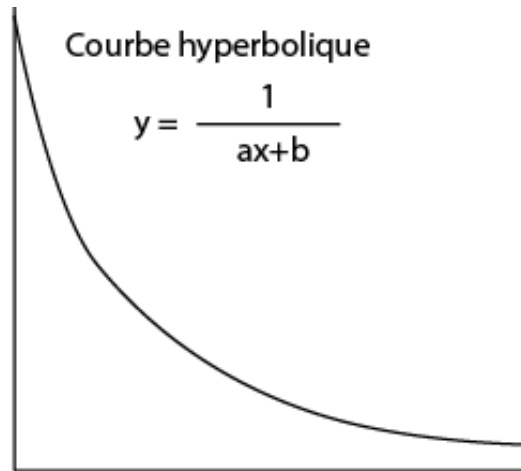
## 2. Analyse de regression – relation exponentielle

---

Extrêmement facile mais attention quand même...!!!

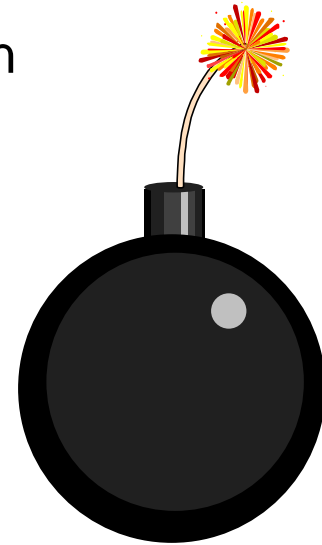
L'ajustement par moindres carrés de la droite  $y = ax + b$  suppose que toutes les mesure  $y_1, \dots, y_n$  soient également incertaines.

## 2. Analyse de regression – Les autres grands modèles

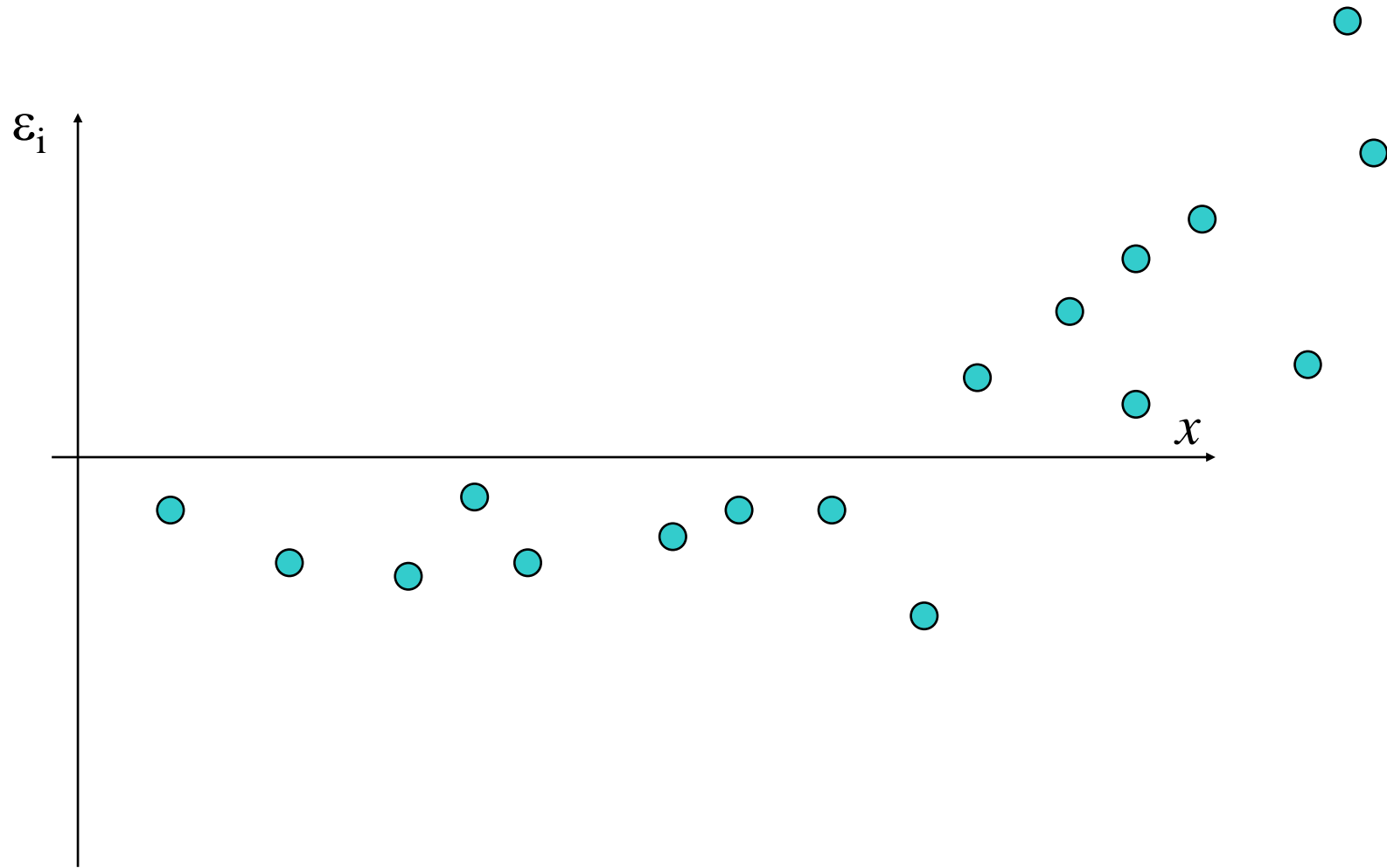


### *Attention*

- Les points isolés ont un effet indésirables sur la régression  
Leur influence doit être testée en les éliminant et en répétant la régression.
- La différence en  $y$  entre un point et la droite de régression est connue sous le nom de résidu.  
La validité de la régression statistique dépend de la distribution des résidus:
  1. Les résidus doivent être normalement distribués
  2. Il ne doit pas y avoir de tendance dans la distribution de variance le long de  $x$ .

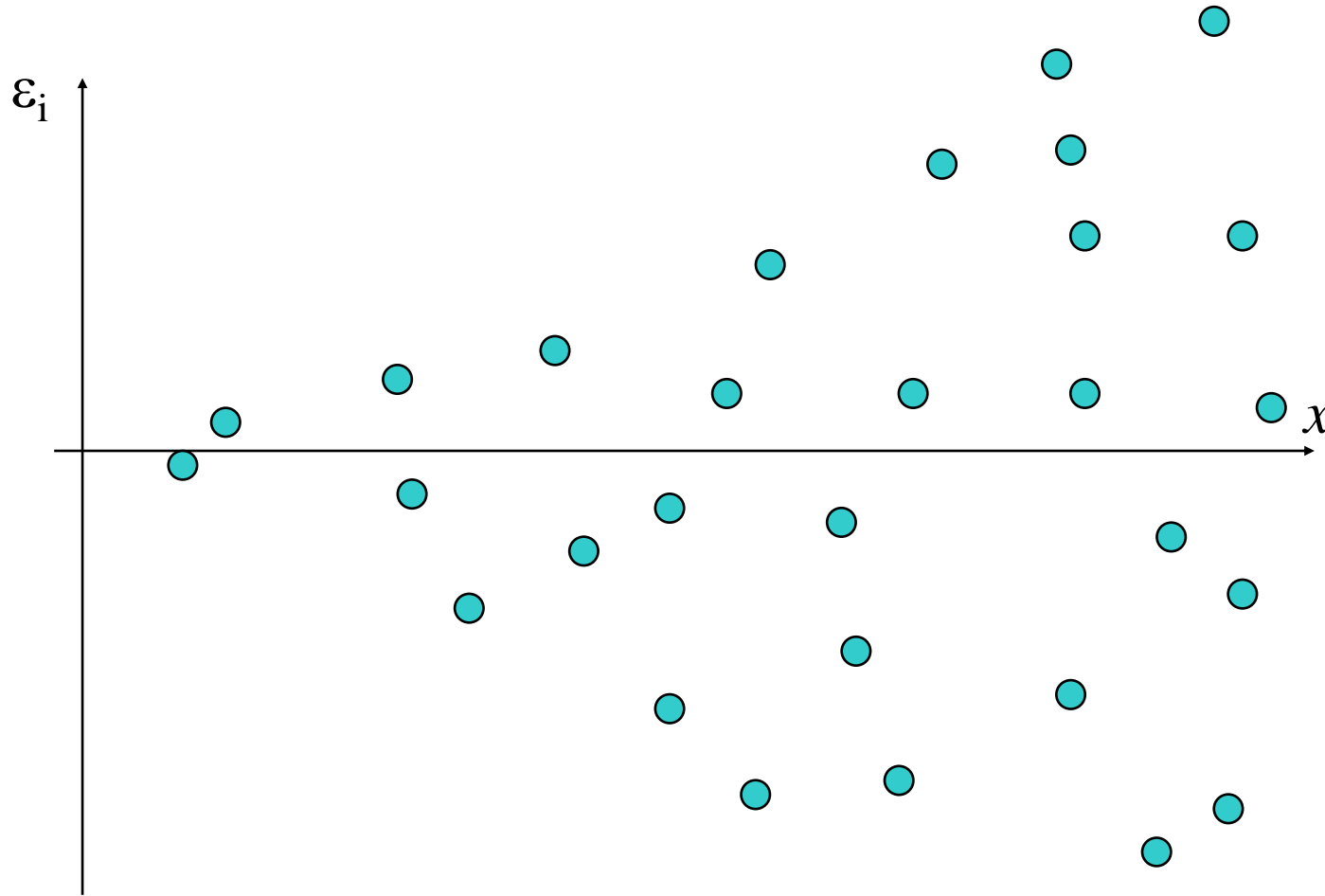


## 2. Analyse de regression – Et les résidus...?

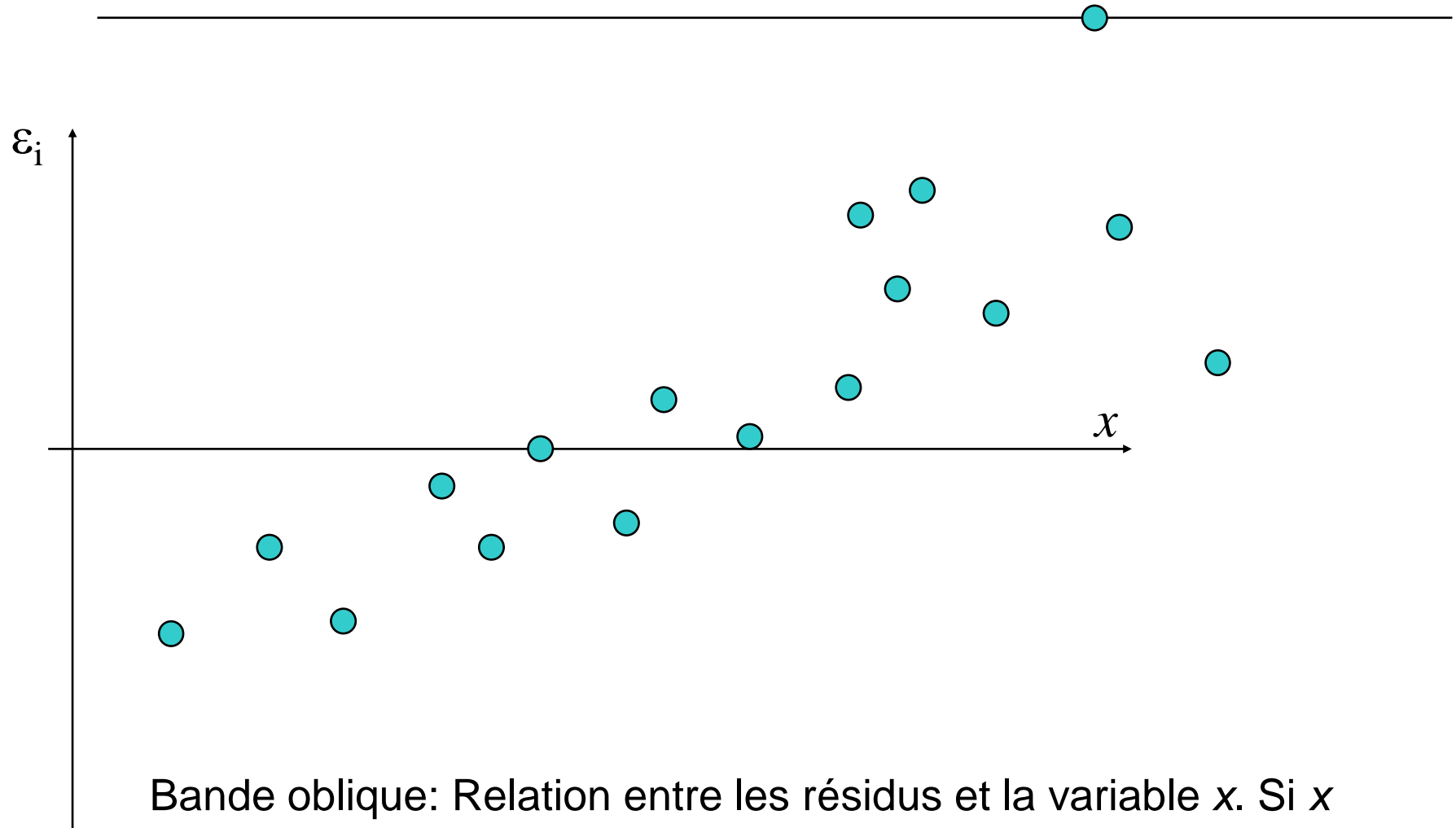


## 2. Analyse de regression – Et les résidus...?

Le fuseau: La variance des résidus n'est pas indépendante des valeurs \_\_\_\_\_ de  $x$ . Des corrections doivent être apportées (courbe log. log p.e.)



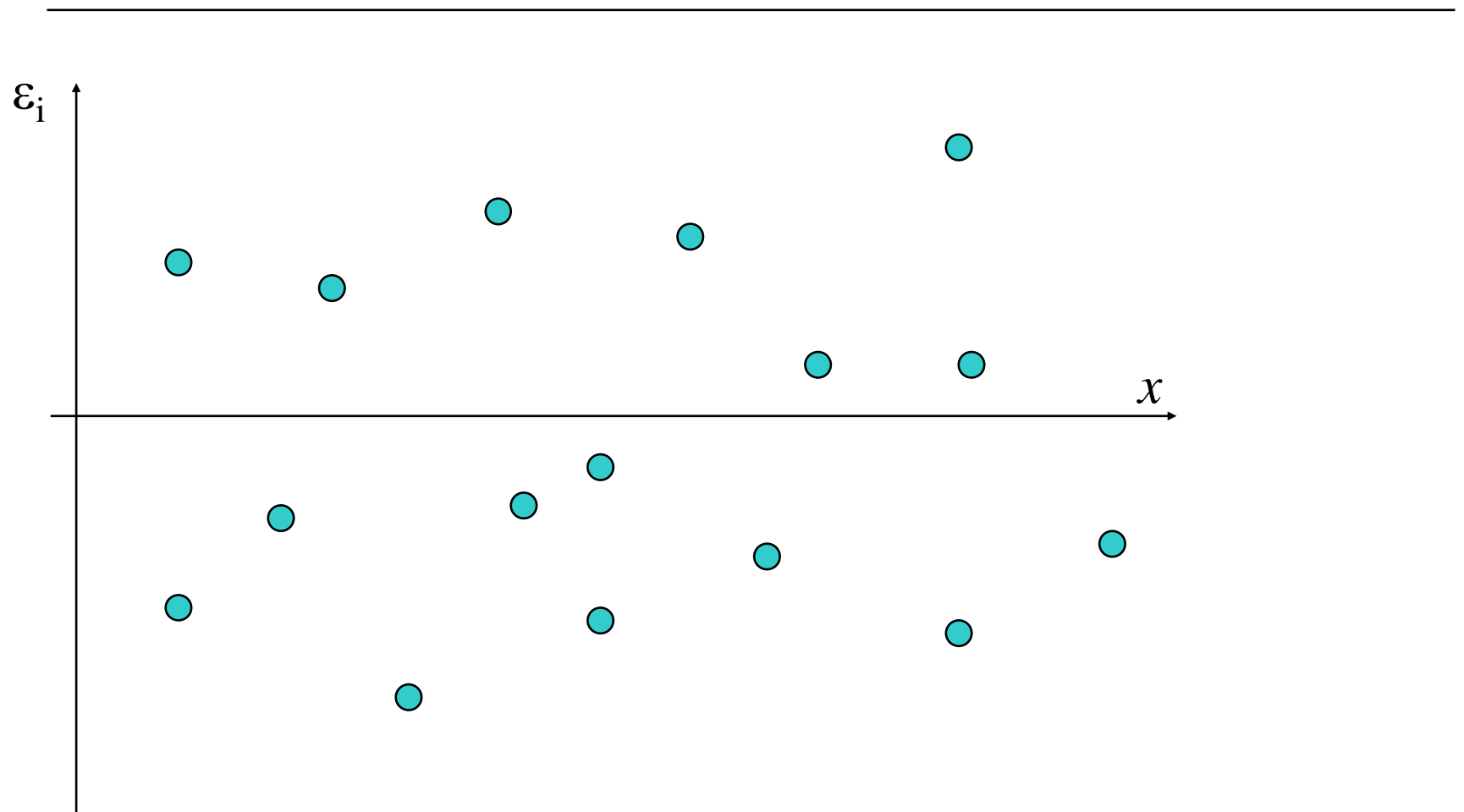
## 2. Analyse de regression – Et les résidus...?



Bande oblique: Relation entre les résidus et la variable  $x$ . Si  $x$  n'est pas dans le modèle, il faudrait l'introduire, ou erreur importante.

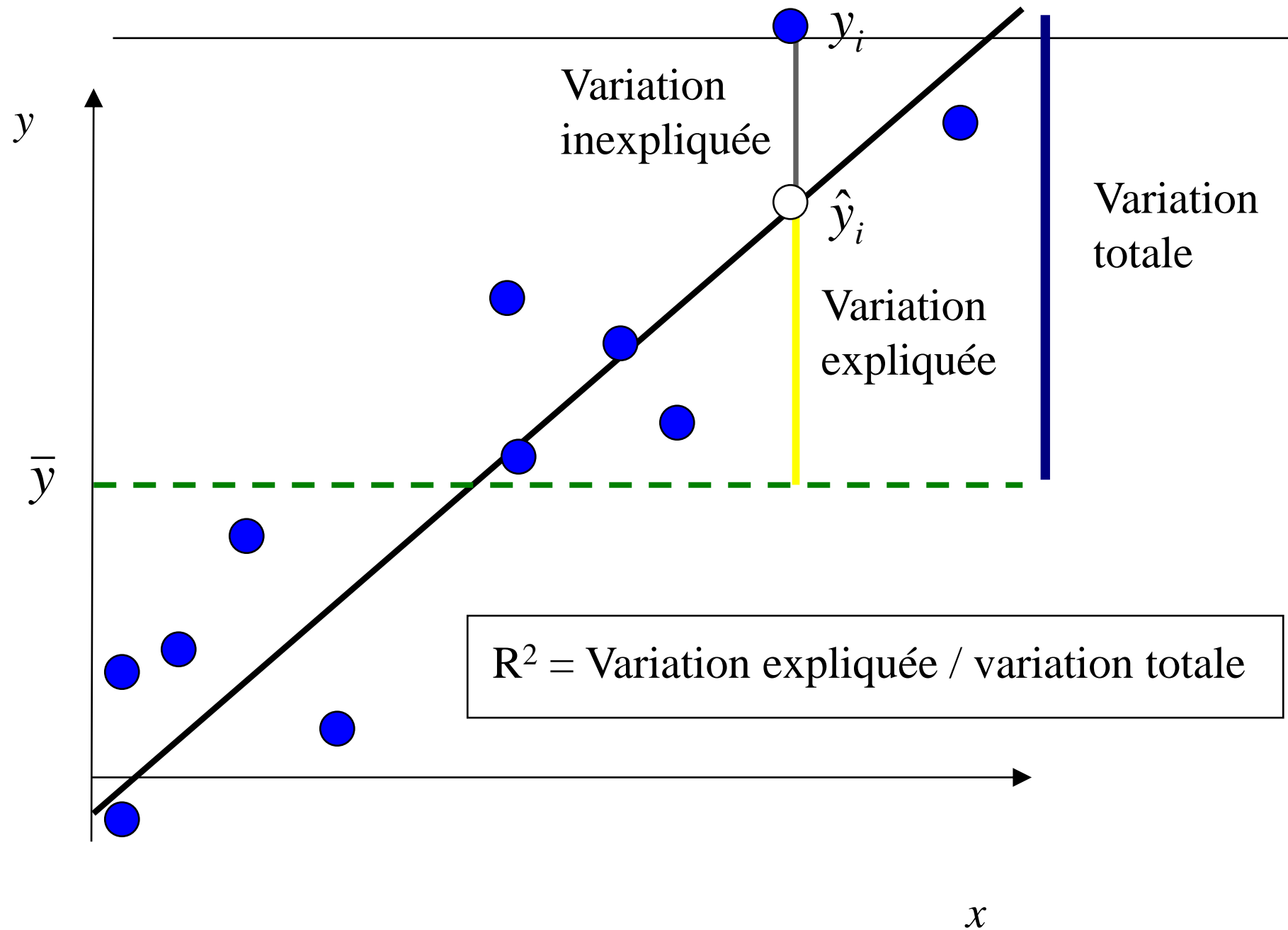


## 2. Analyse de regression – Et les résidus...?



Bande horizontale: les conditions d'application sont suffisamment respectées

## 2. Analyse de regression – Le coefficient de détermination



## 2. Analyse de regression – Le coefficient de détermination

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Somme des carrés  
totale ( $SC_{\text{tot}}$ )

Somme des carrés  
des résidus ( $SC_{\text{res}}$ )

Somme des carrés  
de la régression ( $SC_{\text{reg}}$ )

Variation totale = variation inexpliquée + variation expliquée

$$R^2 = \text{Variation expliquée} / \text{variation totale}$$

$R^2$  est le coefficient de détermination, proportion de la variation de  $y$  qui s'explique par la présence de  $x$ .

Plus  $R^2$  est grand, plus  $SC_{\text{res}}$  est petit.