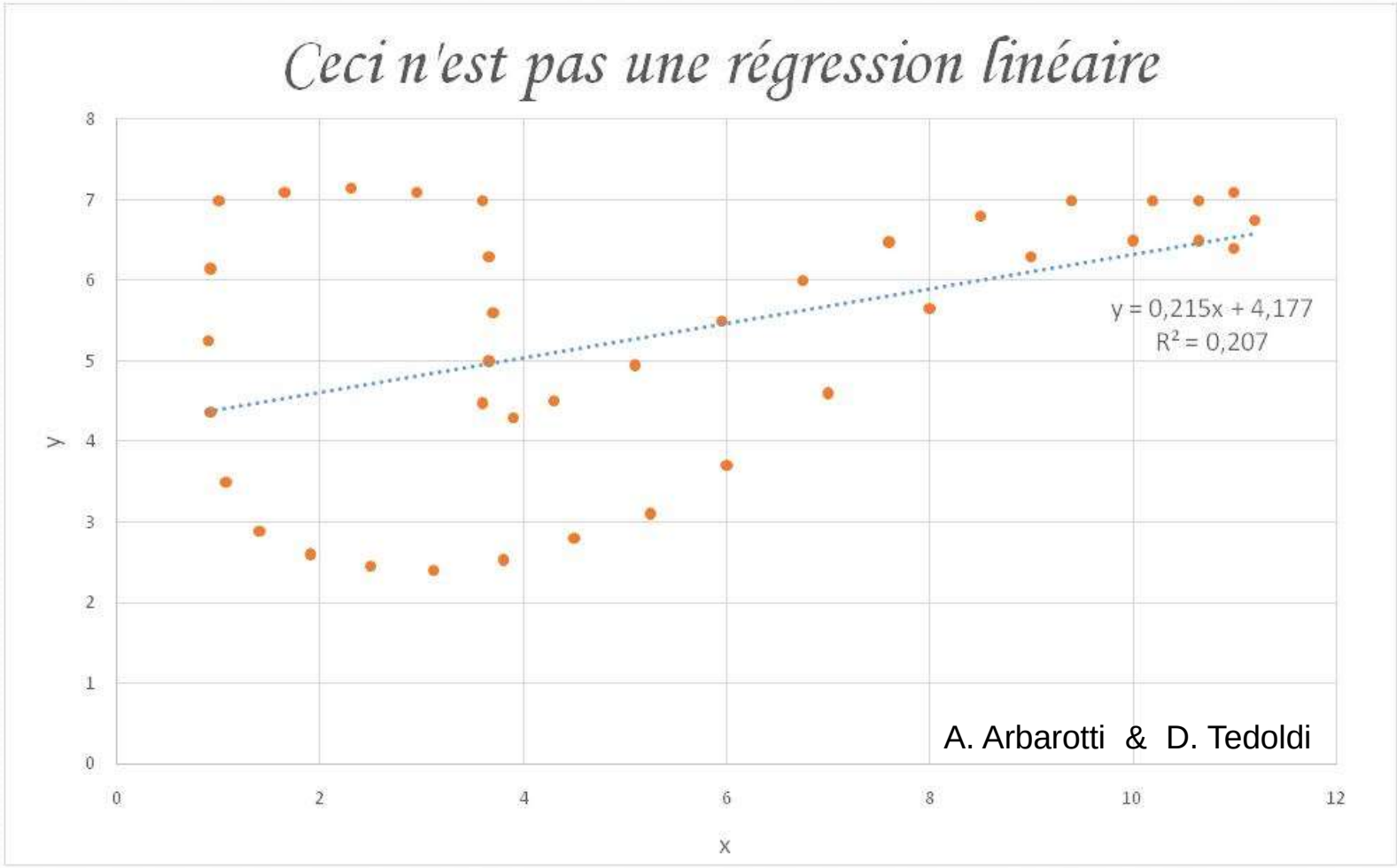
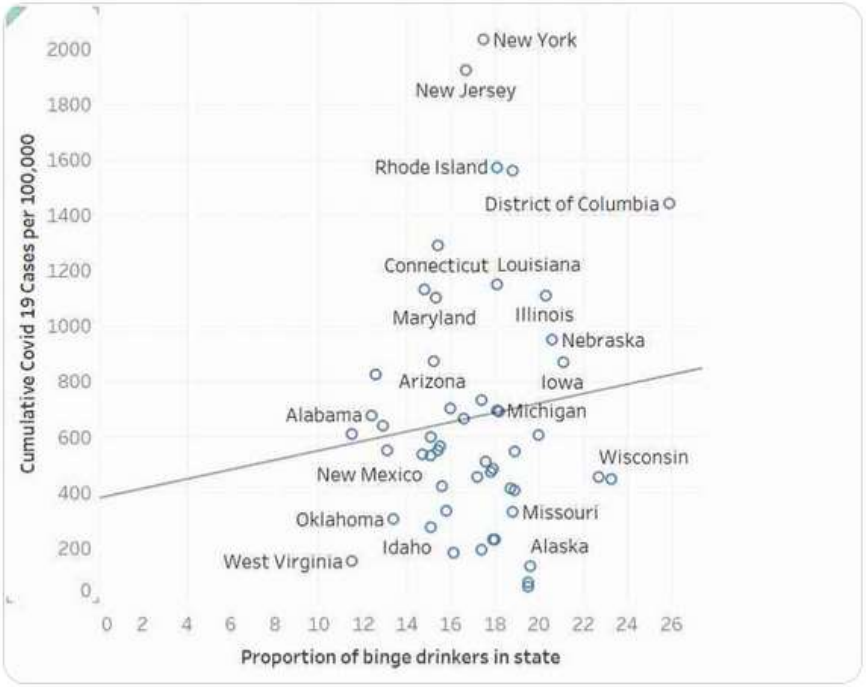
Statistiques
Retour aux fondamentaux



Petit florilège

Amihai Glazer
@AmihaiGlazer

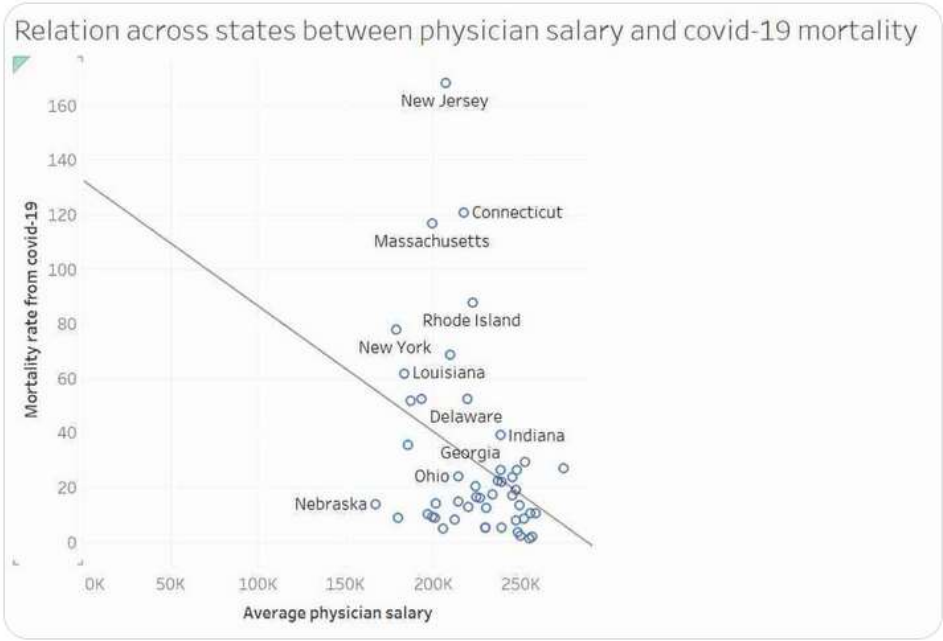
Use prevalence of binge drinking in a state as a proxy for frequency of visits to bars. So some evidence that bar visits are associated with more covid-19 cases. Consistent with the idea that re-opening bars caused problems.



2:28 AM · 4 juil. 2020 · Twitter Web App

Amihai Glazer
@AmihaiGlazer

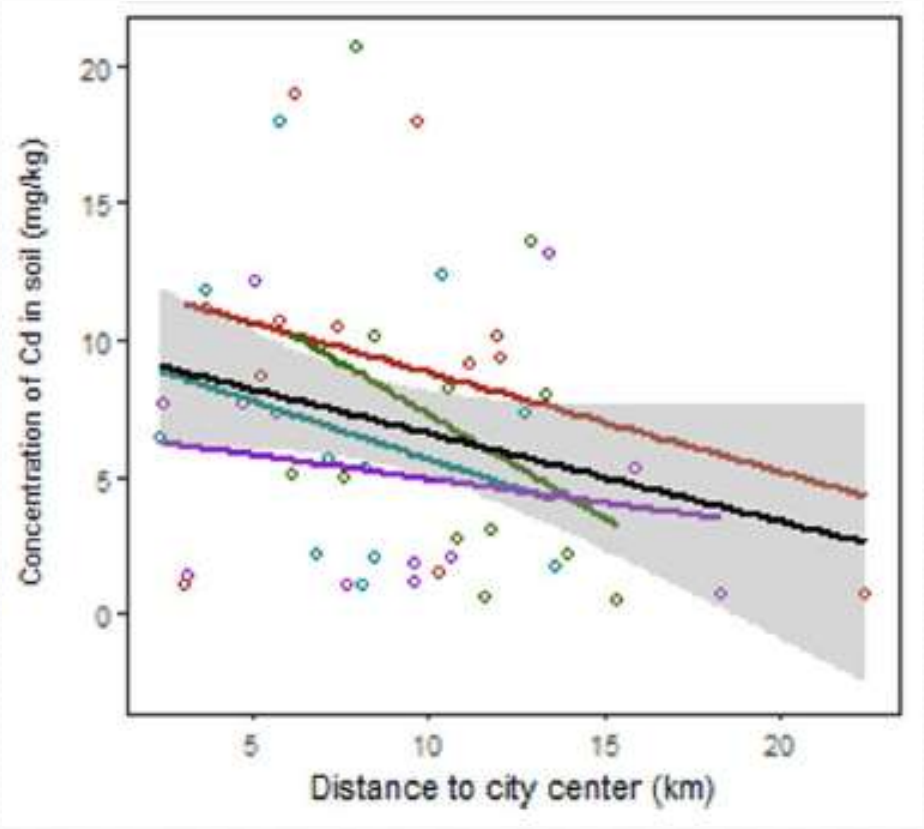
States where physicians are highly paid have lower Covid-19 mortality per capita.

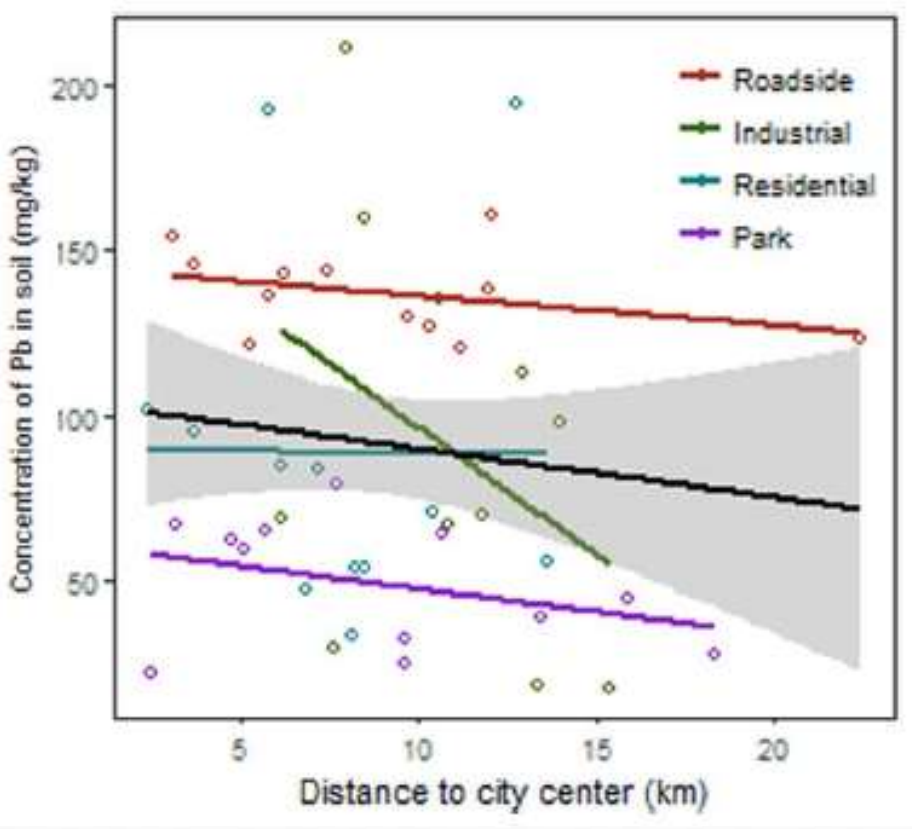


3:03 AM · 30 juin 2020 · Twitter Web App

Twitter (sans trucage)

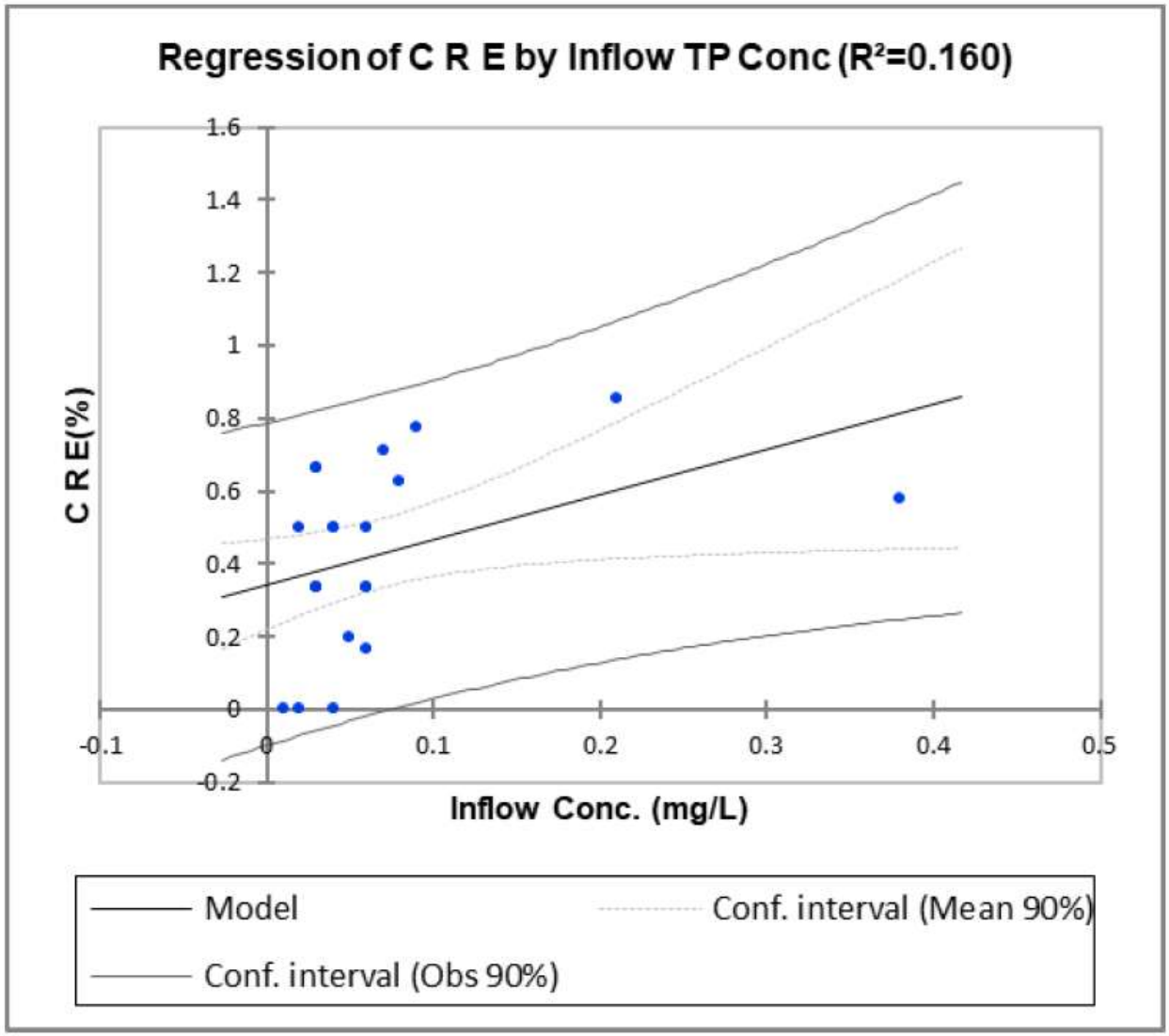
Petit florilège





Zhang et al., 2018, *Nature Scientific Reports*

Petit florilège



Drapper & Hornbuckle, 2018

Petit florilège

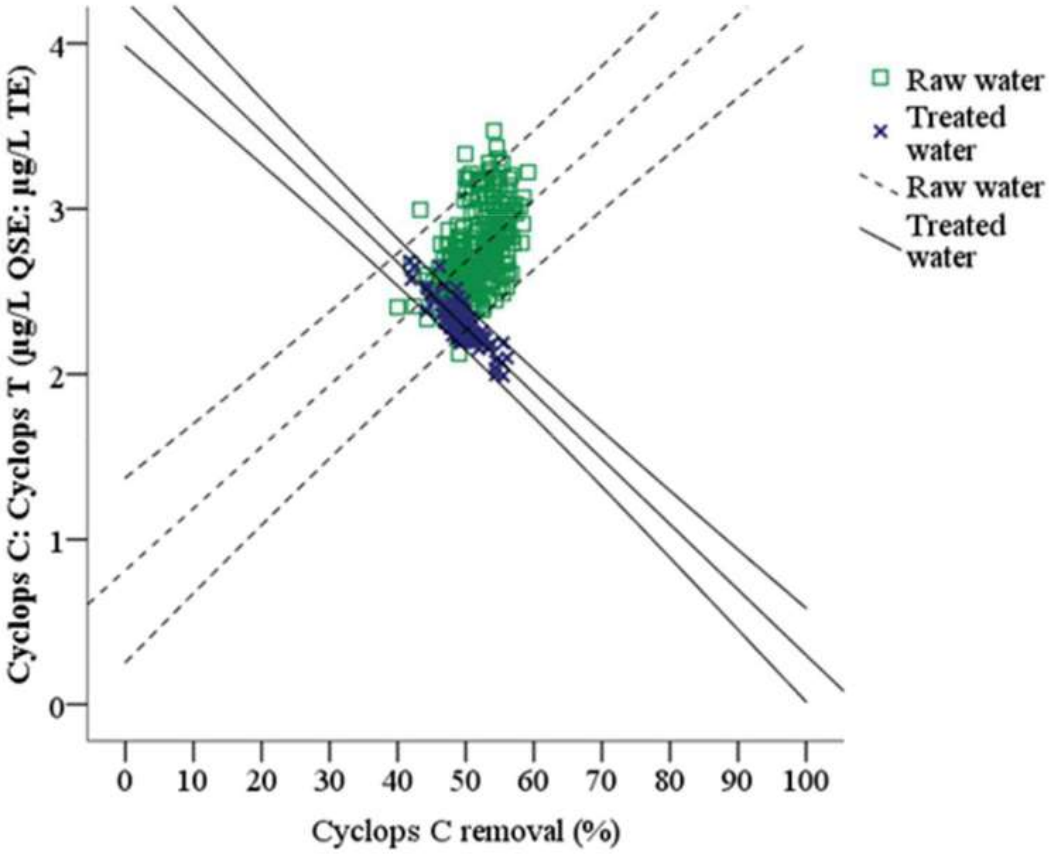


Fig. 8 Correlation between OM character changes in raw treated water and OM removal in (a) Capalaba WTP and (b) Yarra Glen WTP. The lines represent the linear relationship and 95% confidence interval between the Cyclops C to Cyclops T ratio and Cyclops C removal, where the dashed line indicated raw water and the solid line the treated water correlations.

Shutova et al., 2016, *Environmental Science : Water Research and Technology*

Petit florilège

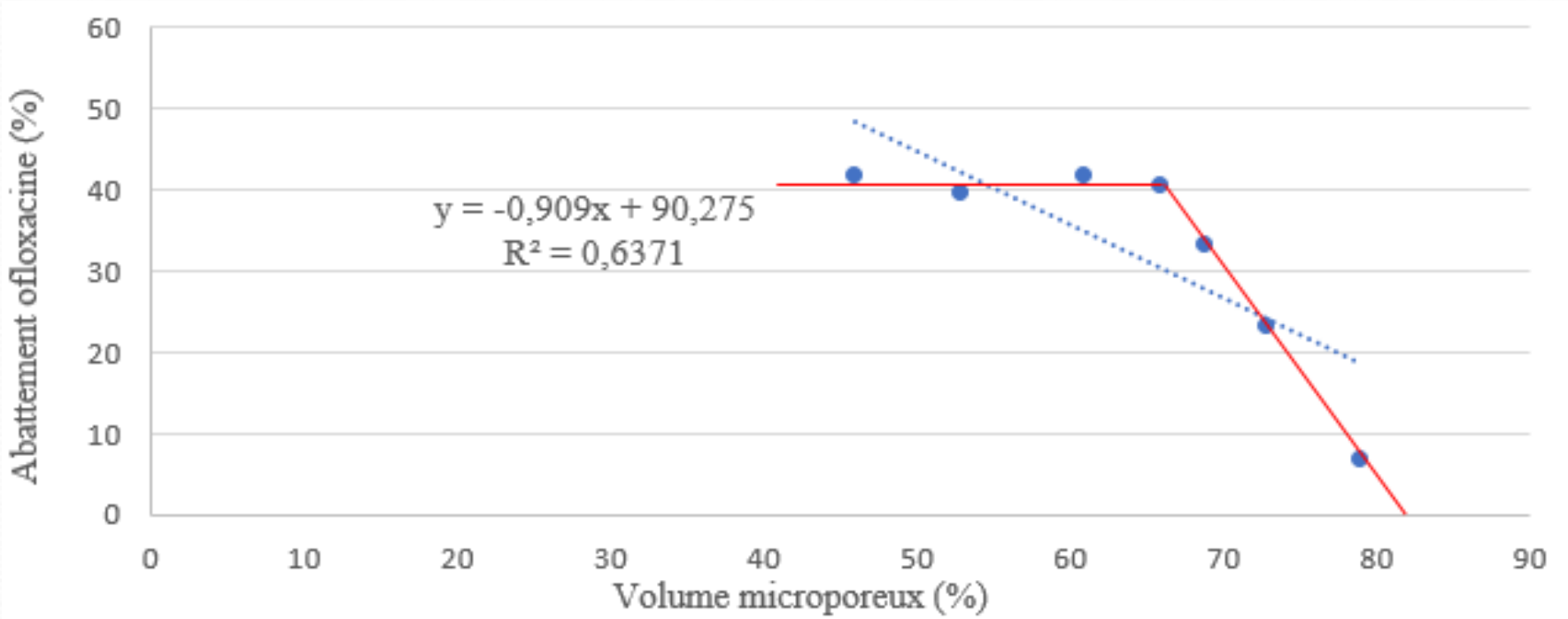
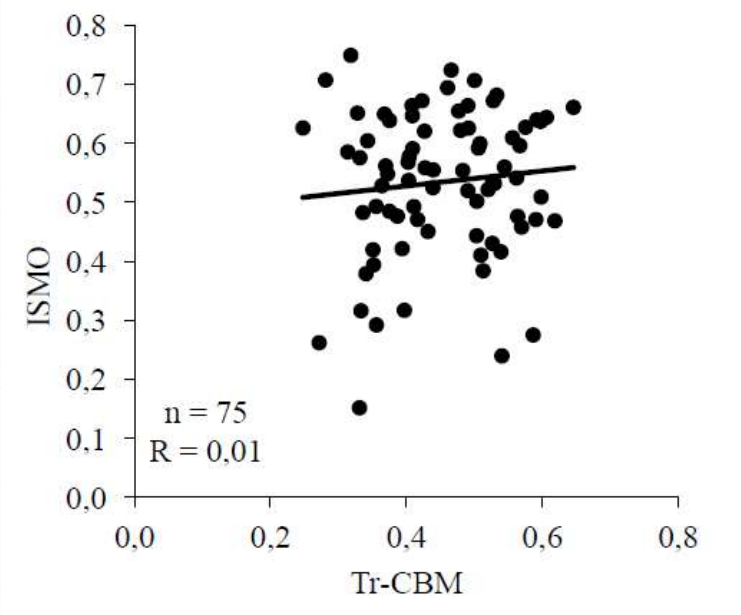


Figure 6 : Abattement de l'ofloxacine (%) en fonction du volume microporeux (%)

Un rapport de stage de M2…

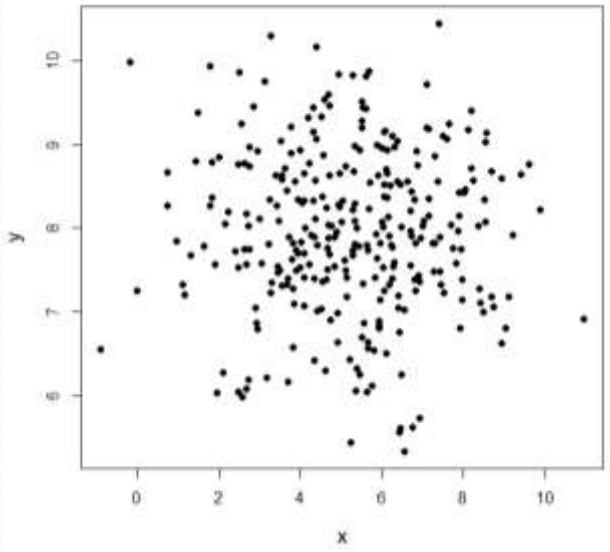
Contenu de la séance

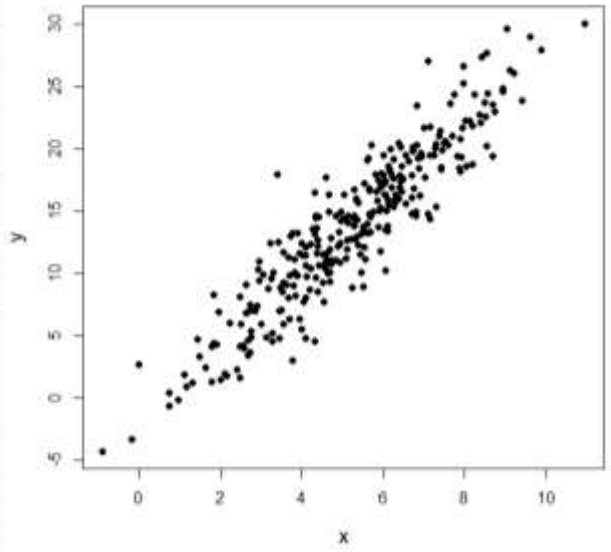
- Qu’est-ce qu’une corrélation ? À quoi ça sert ?
- Vous avez dit « *significatif* » ?
- Pearson ou Spearman ?
- Et la régression linéaire dans tout ça ?

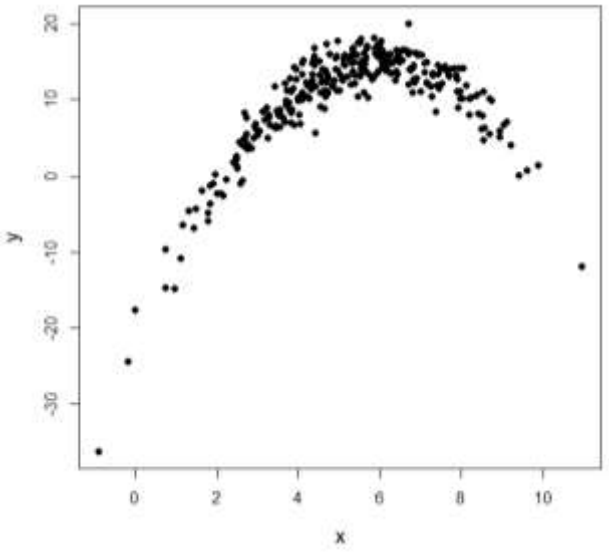


Qu'est-ce qu'une corrélation ?

- Une mesure du « degré de liaison » entre deux variables quantitatives x et y $\longrightarrow$ notion à préciser !
- Existe-t-il une relation (linéaire ou non) entre x et y **qui soit prépondérante devant leurs fluctuations internes** ?







À quoi ça sert ?

- Utiliser x comme mesure indirecte de y
  - Exemple classique : turbidité comme proxy des MES
  - Modélisation empirique = qui découle de l'observation
- Réduire le nombre de dimensions d'un problème…
  - Principe central de l'Analyse en Composantes Principales
- …ou démontrer au contraire la nécessité de considérer deux variables non-corrélées
- Identifier l'existence d'une cause commune aux variations de x et y (« phénomène-source »)

Prudence : corrélation ⇏ causalité

Le Monde

ACTUALITÉS · ÉCONOMIE · VIDÉOS · OPINIONS · CULTURE · M LE MAG · SERVICES

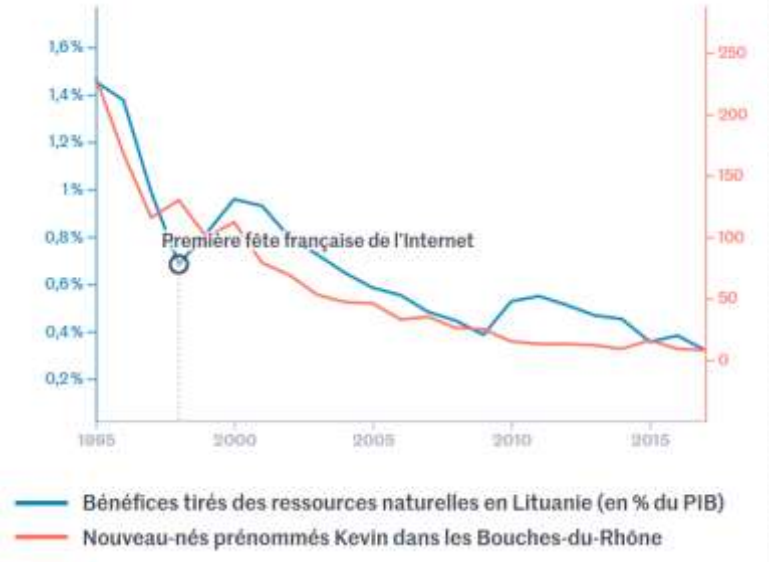
LES DÉCODEURS · DATAVISUALISATION

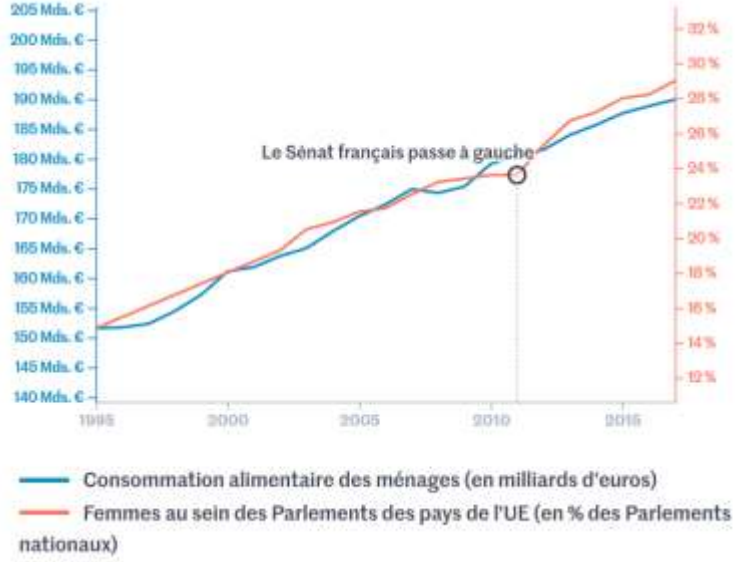
Partage

Corrélation ou causalité ? Brillez en société avec notre générateur aléatoire de comparaisons absurdes

Internet voit parfois émerger des courbes ou des cartes qui prétendent pouvoir expliquer simplement des questions complexes : quelques conseils pour ne pas tomber dans le panneau.

Par Pierre Breteau, Maxime Ferrer et Lucas Baudin · Publié le 02 janvier 2019 à 07h11 · Mis à jour le 06 mars 2019 à 10h28





Mesurer la corrélation entre 2 variables

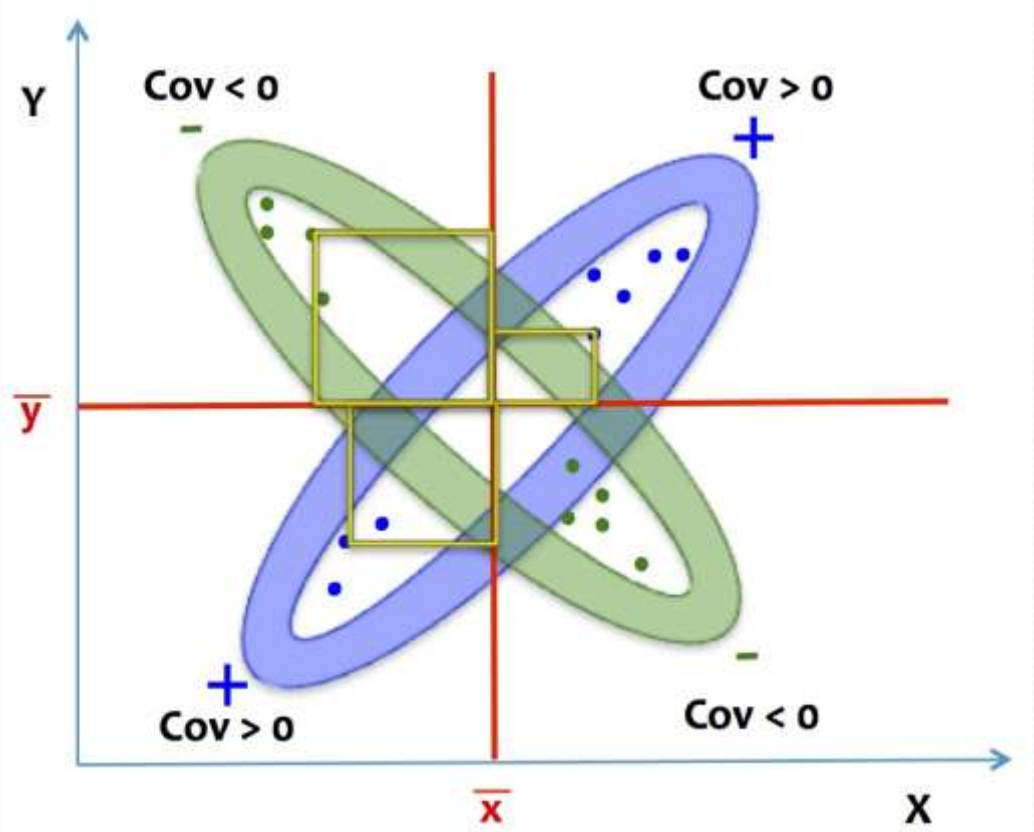
- Que nous dit la covariance ?

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big((X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big)$$

En moyenne…

…quand X est supérieur à son espérance…

…Y tend-il à être supérieur ou inférieur à son espérance ?



Hugues Michel, 2015

Mesurer la corrélation entre 2 variables

- Que nous dit la covariance ?

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big((X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big)$$

En moyenne…

…quand X est supérieur à son espérance…

…Y tend-il à être supérieur ou inférieur à son espérance ?

- $\mathrm{Cov}(X,Y) > 0$ si X et Y ont tendance à fluctuer dans le même sens
- inversement si $\mathrm{Cov}(X,Y) < 0$
- $\mathrm{Cov}(X,Y) = 0 \Rightarrow$ les termes + et – se compensent $\not\Rightarrow$ X et Y sont indépendants (en général)

Mesurer la corrélation entre 2 variables

- Quel est le problème ?
  - Comment interpréter la valeur de $\mathrm{Cov}(X, Y)$?
    - Gamme de valeurs prises par la covariance :

$$-\infty < \mathrm{Cov}(X, Y) < +\infty$$

    - …et ensuite ? La liaison entre X et Y est-elle « forte » ou non ?
  - Comment comparer des covariances ?
    - Dimension de $\mathrm{Cov}(X, Y)$ = produit des dimensions de X et Y
    - $[\mathrm{Cov}(\mathrm{Turbidité}, \mathrm{MES})]$ = $\mathrm{NTU.mg.L^{-1}}$ (???)
  - Comparaison *en relatif*, pour des *couples identiques*

- **Besoin d'un indicateur normalisé : la *corrélation***

Mesurer la corrélation entre 2 variables

- Le coefficient de corrélation **linéaire** de Pearson

$$r_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \, \sigma_Y}$$

- Cette fois-ci, $-1 \leq r_{XY} \leq 1$ (sans unité)
  - Démo : Cauchy-Schwarz, en notant que l'application $(X - \mathbb{E}[X], Y - \mathbb{E}[Y]) \longmapsto \mathrm{Cov}(X, Y)$ définit un produit scalaire
- $|r_{XY}| = 1 \Longrightarrow X$ et Y sont liés par une relation linéaire
  - Démo : cas d'égalité dans Cauchy-Schwarz
- $r_{XY} = 0 \Longrightarrow$ pas de relation **linéaire** apparente

Mesurer la corrélation entre 2 variables

- On progresse, mais… où placer le seuil entre 0 et 1 ?
- À partir de quand une corrélation est-elle *significative* ?
  - Test d'hypothèse !
  - (glissement vers les statistiques inférentielles)

$$\mathcal{H}_0 : r_{XY} = 0$$

$$\mathcal{H}_1 : r_{XY} \neq 0$$

- Question posée dans un test stat : les observations disponibles sont-elles compatibles avec $\mathcal{H}_0$?

Rappel : principe d'un test statistique

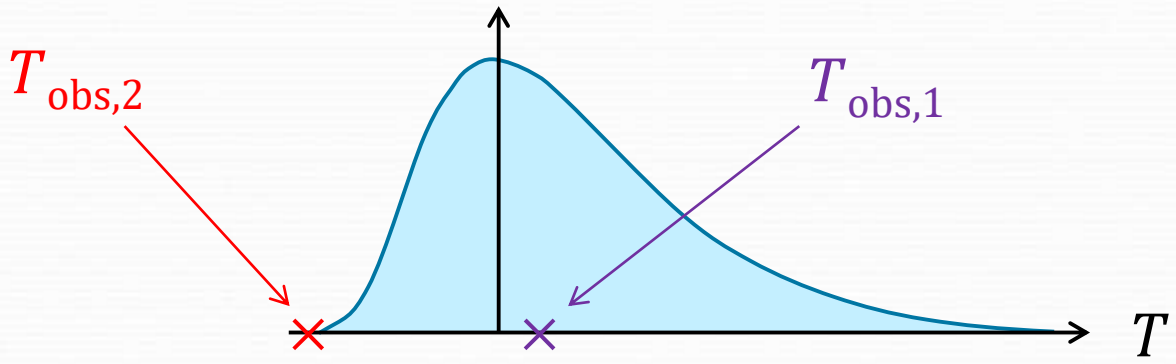
- « Raisonnement par l'absurde probabiliste »
- On suppose que $\mathcal{H}_0$ est vraie
- …
- … (calculs)
- …
- Contradiction ?
  - Non pas une contradiction mathématique (« 3 = 2 »)
  - Mais une valeur (très) peu probable
  - Dans ce cas, on peut choisir de rejeter $\mathcal{H}_0$

Rappel : principe d'un test statistique

- On calcule une « statistique de test » en fonction des observations disponibles :

$$T = f(x_1, x_2, \dots, x_n)$$

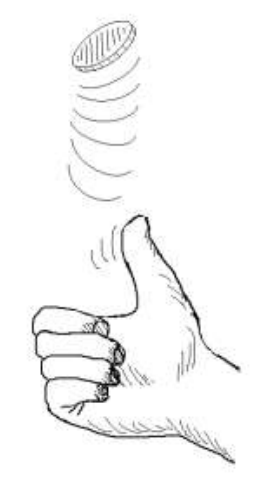
- **Si $\mathcal{H}_0$ est vraie** (+ éventuellement d'autres hypothèses) alors la distribution **théorique** de T est connue



- Dans ce cas, la valeur observée est-elle *vraisemblable* ?

Rappel : principe d'un test statistique

- Un exemple simple : « ma pièce est-elle équilibrée ? »

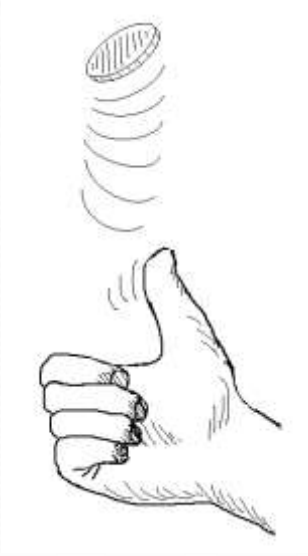


× 20 ⇒ 15 « pile », 5 « face »

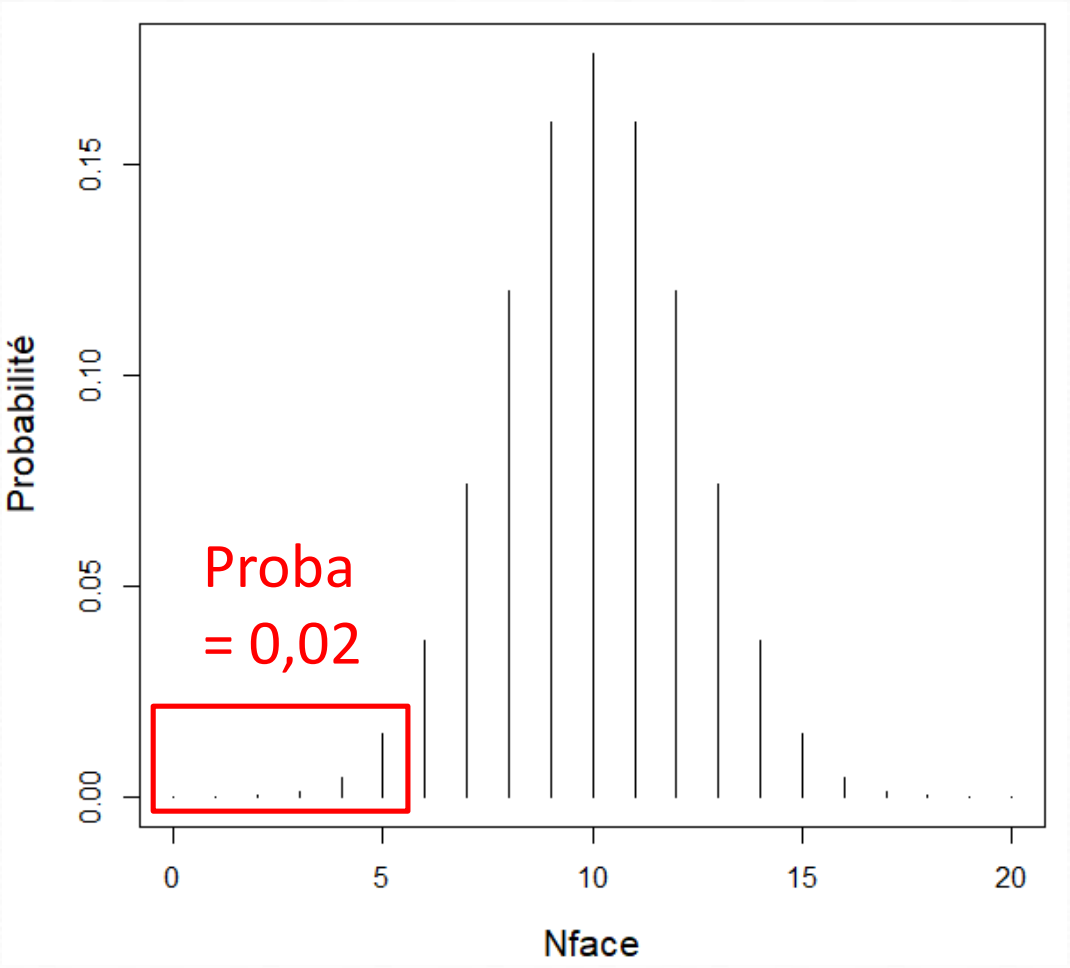
- **Si la pièce était équilibrée** ($\mathcal{H}_0 : \mathbb{P}(\text{face}) = 0.5$), un tel résultat serait-il vraisemblable (= pourrait-il être dû simplement au hasard) ?
- **Distribution théorique** : $N_{\text{face}} \sim \mathcal{B}(20, 0.5)$

Rappel : principe d'un test statistique

- **Distribution théorique** : $N_{\text{face}} \sim \mathcal{B}(20, 0.5)$



$\times$ 20

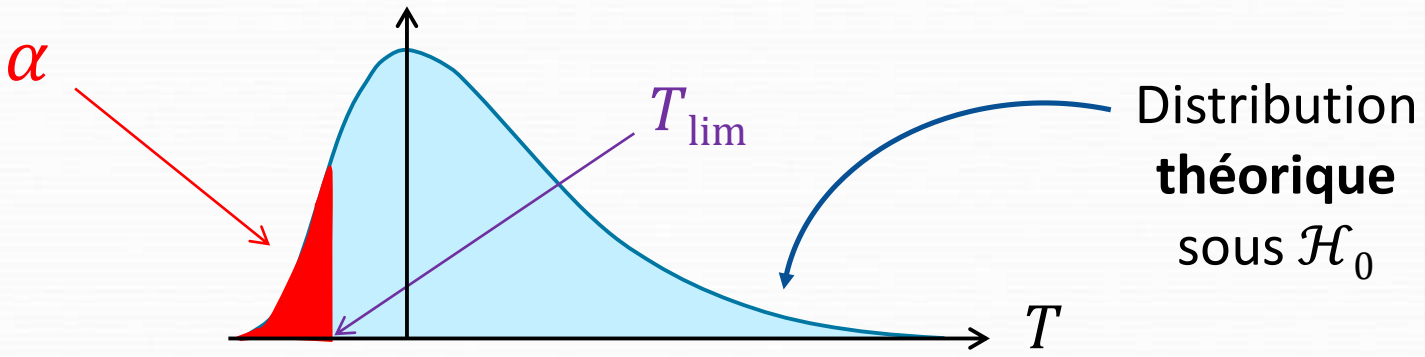


$\Longrightarrow$ On rejette $\mathcal{H}_0$?

Rappel : principe d'un test statistique

- **Prise de décision** : première façon de procéder
  - Se donner un « seuil de signification » α et déterminer la valeur T_{lim} tel que :

$$\mathbb{P}(T < T_{\text{lim}} \mid \mathcal{H}_0 \text{ vraie}) = \alpha$$

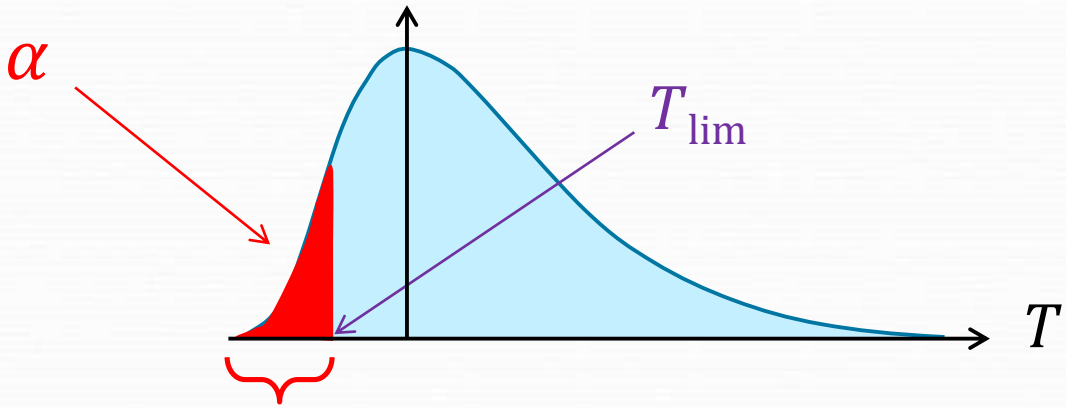


  - Autres versions :
    - Unilatéral à droite : $\mathbb{P}(T > T_{\text{lim}} \mid \mathcal{H}_0 \text{ vraie})$
    - Bilatéral symétrique : $\mathbb{P}(|T| > T_{\text{lim}} \mid \mathcal{H}_0 \text{ vraie})$

Rappel : principe d'un test statistique

- **Prise de décision** : première façon de procéder

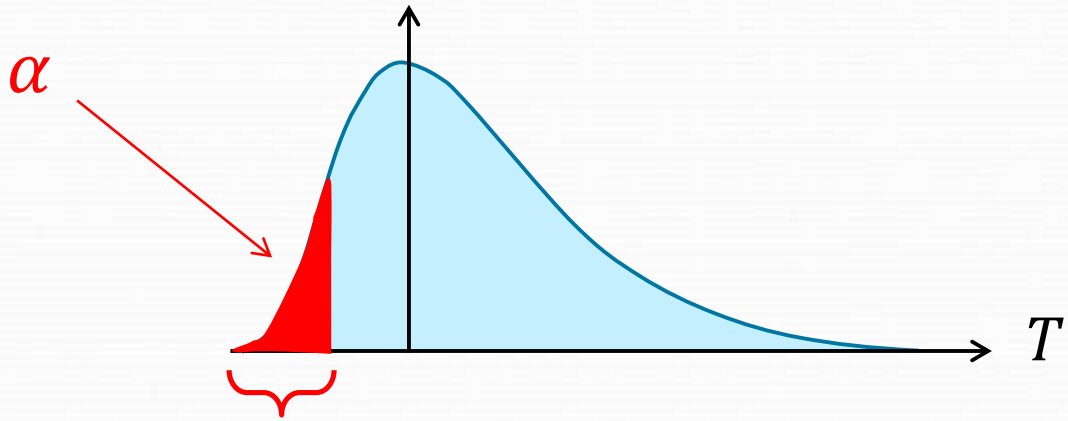
$$\mathbb{P}(T < T_{\text{lim}} \mid \mathcal{H}_0 \text{ vraie}) = \alpha$$



- La région $\{ T_{\text{obs}} < T_{\text{lim}} \}$, ou $\{ T_{\text{obs}} > T_{\text{lim}} \}$ ou $\{ |T_{\text{obs}}| > T_{\text{lim}} \}$ ou $\{ \ldots \}$ selon les cas, est appelée **région de rejet**
- La valeur de T_{obs} est jugée « trop extrême » pour être compatible avec l'hypothèse $\mathcal{H}_0$

Rappel : principe d'un test statistique

- **Prise de décision** : première façon de procéder
- **Quelle valeur donner à α ?**
  - Que représente α ? **Si $\mathcal{H}_0$ est vraie**, la distribution de T est



  - α est la probabilité d'occurrence de la région de rejet sous $\mathcal{H}_0$, *i.e.* la **probabilité de rejeter à tort** l'hypothèse nulle

Rappel : principe d'un test statistique

- **Exemple :** Test médical, $\mathcal{H}_0$: « Le patient est malade »

	$\mathcal{H}_0$ vraie	$\mathcal{H}_0$ fausse
On accepte $\mathcal{H}_0$	OK	On soigne un patient alors qu'il n'est pas malade
On rejette $\mathcal{H}_0$	On renvoie un patient chez lui alors qu'il est malade	OK

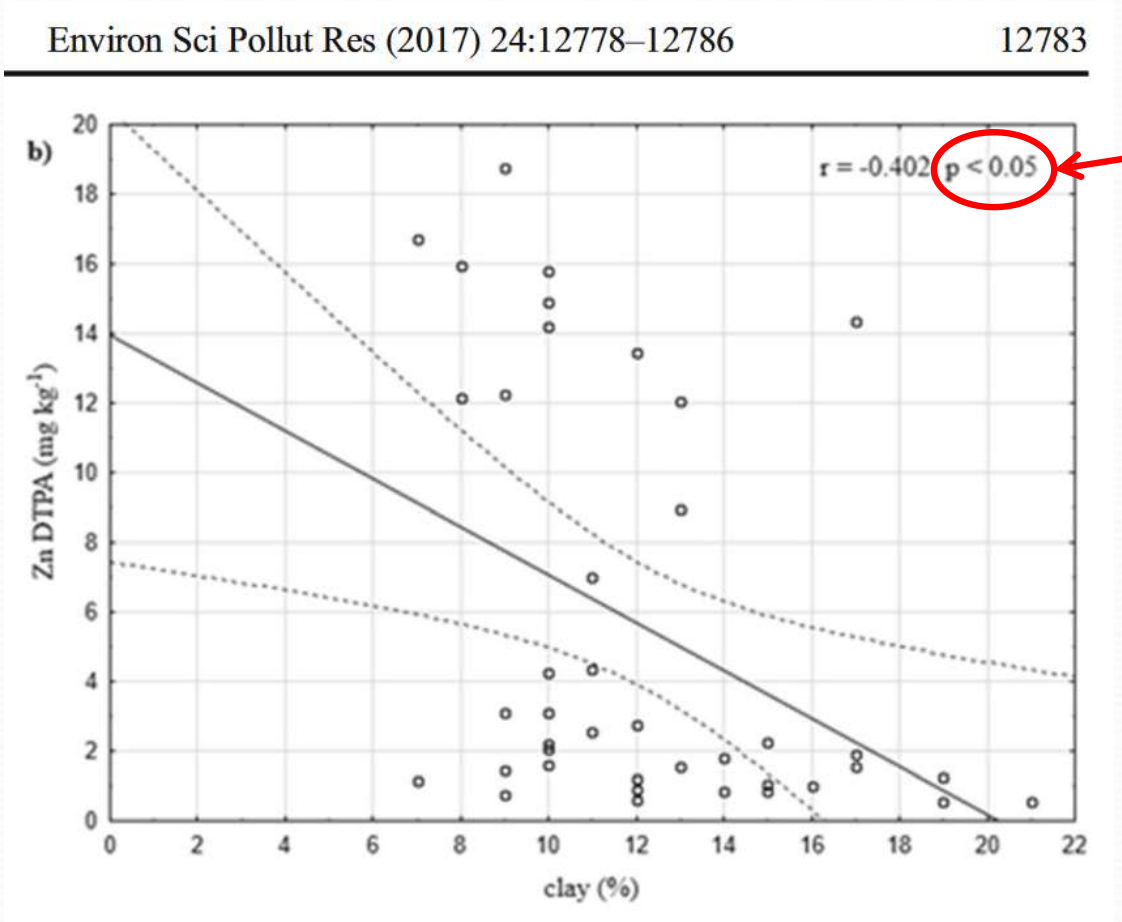
Risque β

Risque α

- Avant de vous demander « quelle valeur pour α ? » (ou pire, de prendre une valeur générique sans réfléchir), demandez-vous : « **quel est l'enjeu** » ?
- **NB** : la définition de $\mathcal{H}_0$ n'est pas neutre !

Rappel : principe d'un test statistique

- **Prise de décision** : seconde façon de procéder
  - ➢ Calculer **la p-valeur** du test (quésaco ?)



Ils sont contents

Rozanski et al., 2017

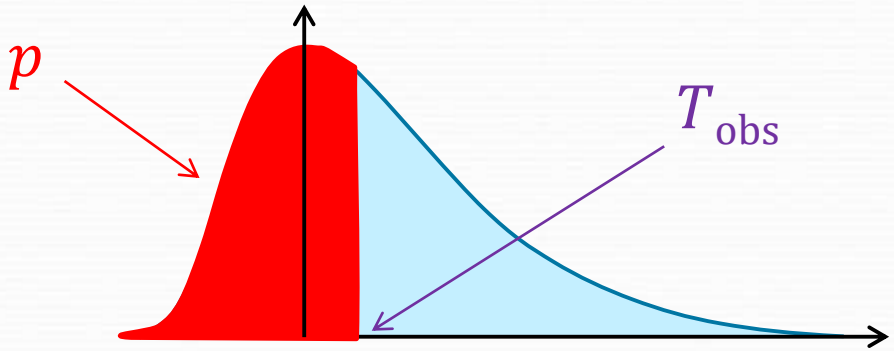
Rappel : principe d'un test statistique

- **Prise de décision** : seconde façon de procéder
  - Calculer **la p-valeur** du test (quésaco ?)

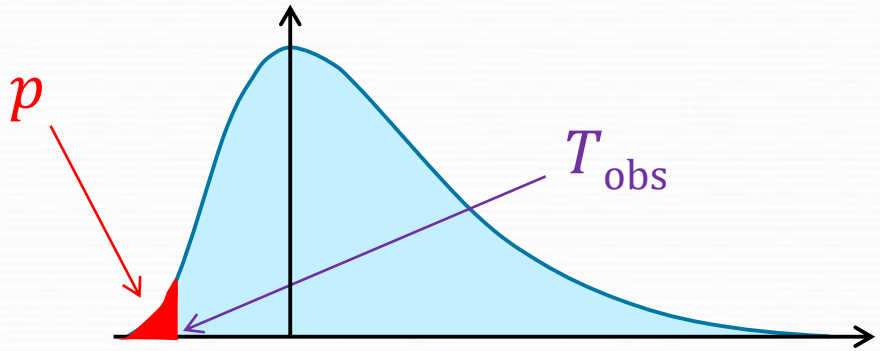
$$p - \text{valeur} = f(T_{\text{obs}}) = \mathbb{P}(T < T_{\text{obs}} \mid \mathcal{H}_0 \text{ vraie})$$

  - (ou toutes les autres versions : unilatérale à droite, etc.)
  - Probabilité de faire « pareil ou encore pire » que la valeur observée, si $\mathcal{H}_0$ est vraie

RAS



Peu vraisemblable (?)



Retour au problème initial

- **La corrélation est-elle significative ?**

$$\mathcal{H}_0 : r_{XY} = 0$$

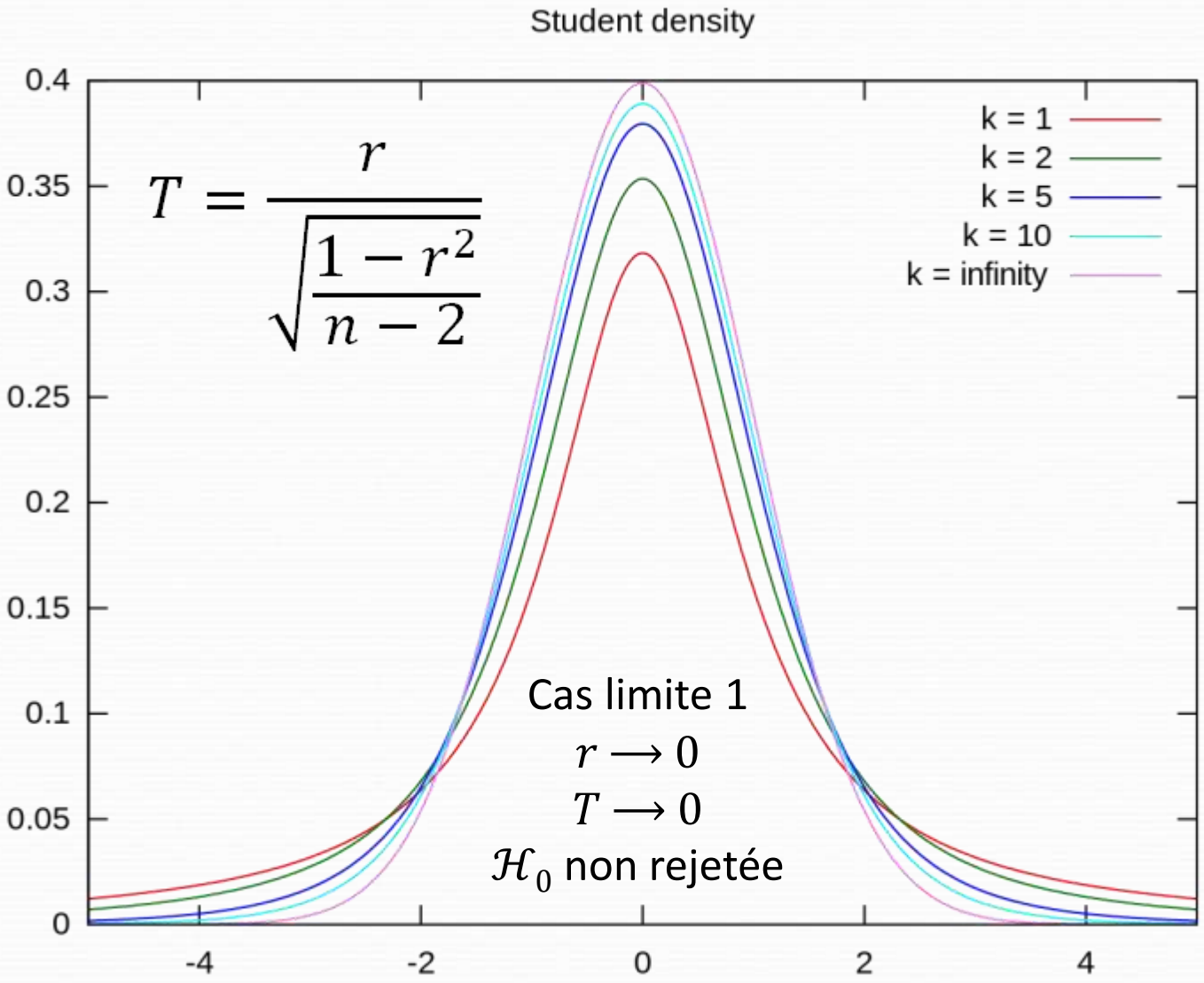
$$\mathcal{H}_1 : r_{XY} \neq 0$$

- **Ai-je une évidence suffisante pour rejeter $\mathcal{H}_0$?**
- On pose :

$$T = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

- Si $\mathcal{H}_0$ est vraie, et si X et Y suivent une loi normale bivariée, alors T suit une loi de Student à $n - 2$ d.d.l.

Retour au problème initial



$$T = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Cas limite 1

$r \longrightarrow 0$

$T \longrightarrow 0$

$\mathcal{H}_0$ non rejetée

Cas limite 2

$r \longrightarrow \pm 1$

$T \longrightarrow \infty$

$\mathcal{H}_0$ rejetée

Matrice des corrélations : interprétation

Table 3
Pearson correlation coefficients and the corresponding scatter plots for soil trace element concentrations (mg/kg) at soil depth ≤5 cm and road distance <5 m, pH (−), LOI (%), and AADT (vehicles/day). Bold type indicates a statistical significant correlation.

	pH	LOI	Cd	Cr	Cu	Ni	Pb	Zn	AADT
pH									
LOI	-0.13								
Cd	-0.25	0.01							
Cr	-0.26	***0.44****	***0.54****						
Cu	0.28	0.01	0.28	-0.04					
Ni	-0.35	0.07	***0.68*****	***0.83*****	0.09				
Pb	***0.32****	0.18	0.26	0.11	***0.62*****	0.15			
Zn	0.16	0.13	***0.36****	0.21	***0.49*****	0.32	***0.69*****		
AADT	-0.30	-0.08	***0.61*****	***0.62*****	***0.39****	***0.58****	0.11	***0.30****	

*P 0.05, statistical significance of the correlations.
**P 0.01, strong statistical significance of the correlations.
***P 0.001, very strong statistical significance of the correlations.

Incohérence (ce n'est pas la seule…)

Werkenthin et al., 2014

À vous : commentaires ?

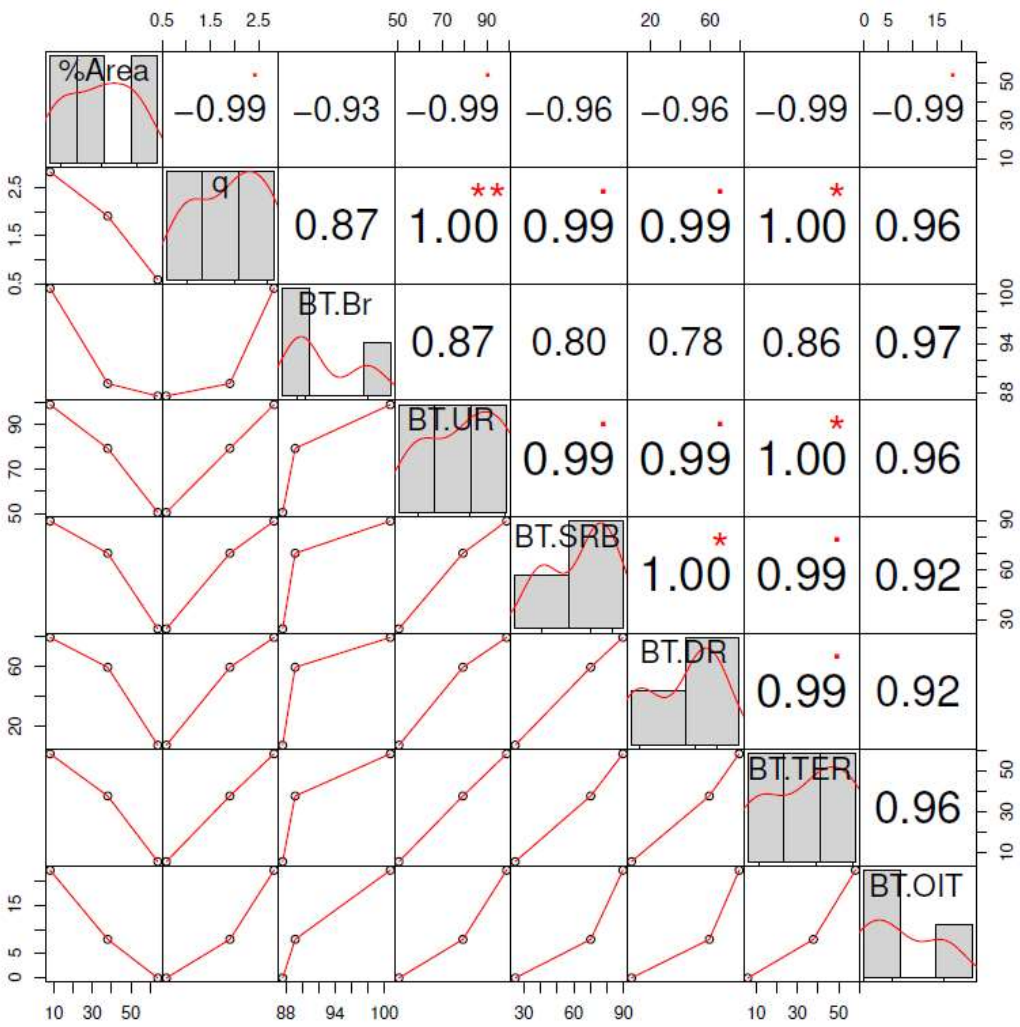
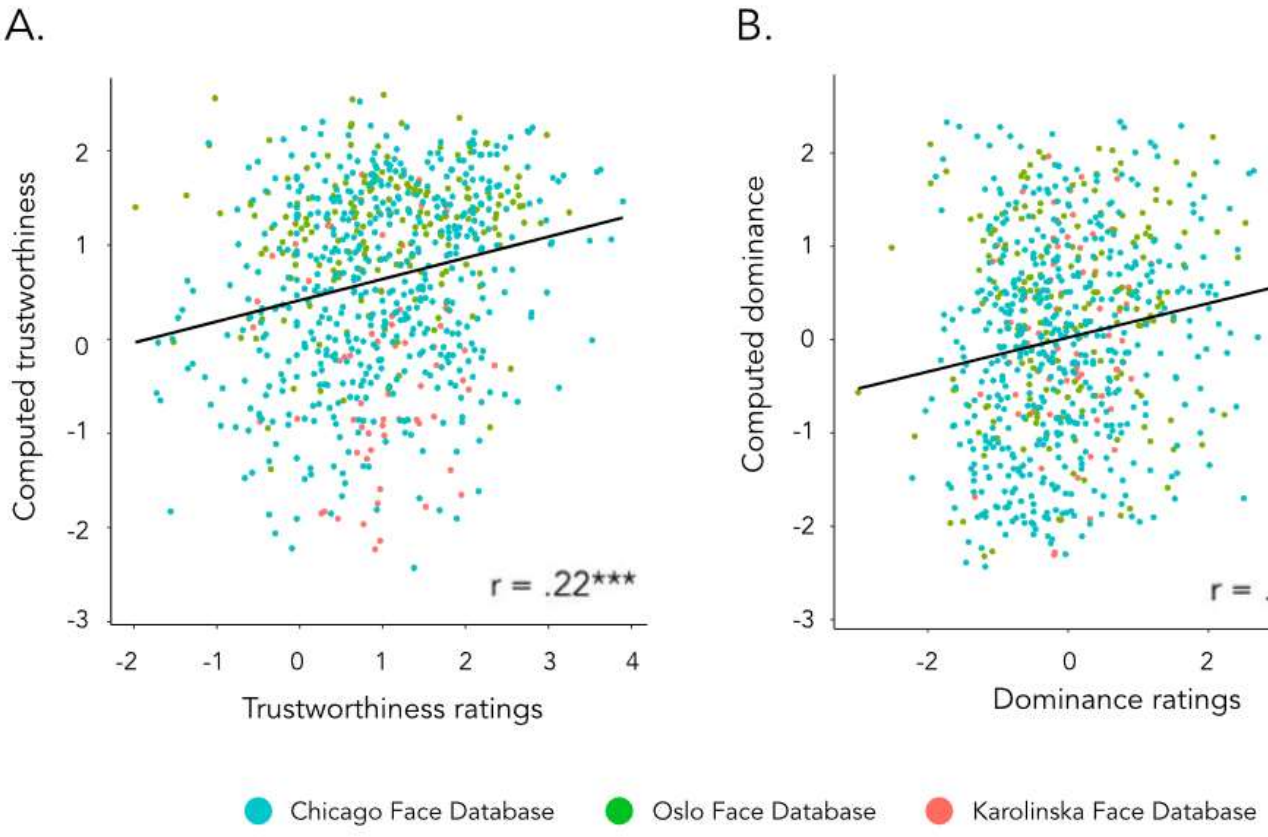


Figure 11. Pearson correlation coefficients for the percentage of the brilliant blue stained area of the soil cut surface of the soil columns and the breakthrough maxima of tracers and biocides (figure was created with the R function "chart.Correlation").

Bork et al., soumis à *Nature Scientific Reports*

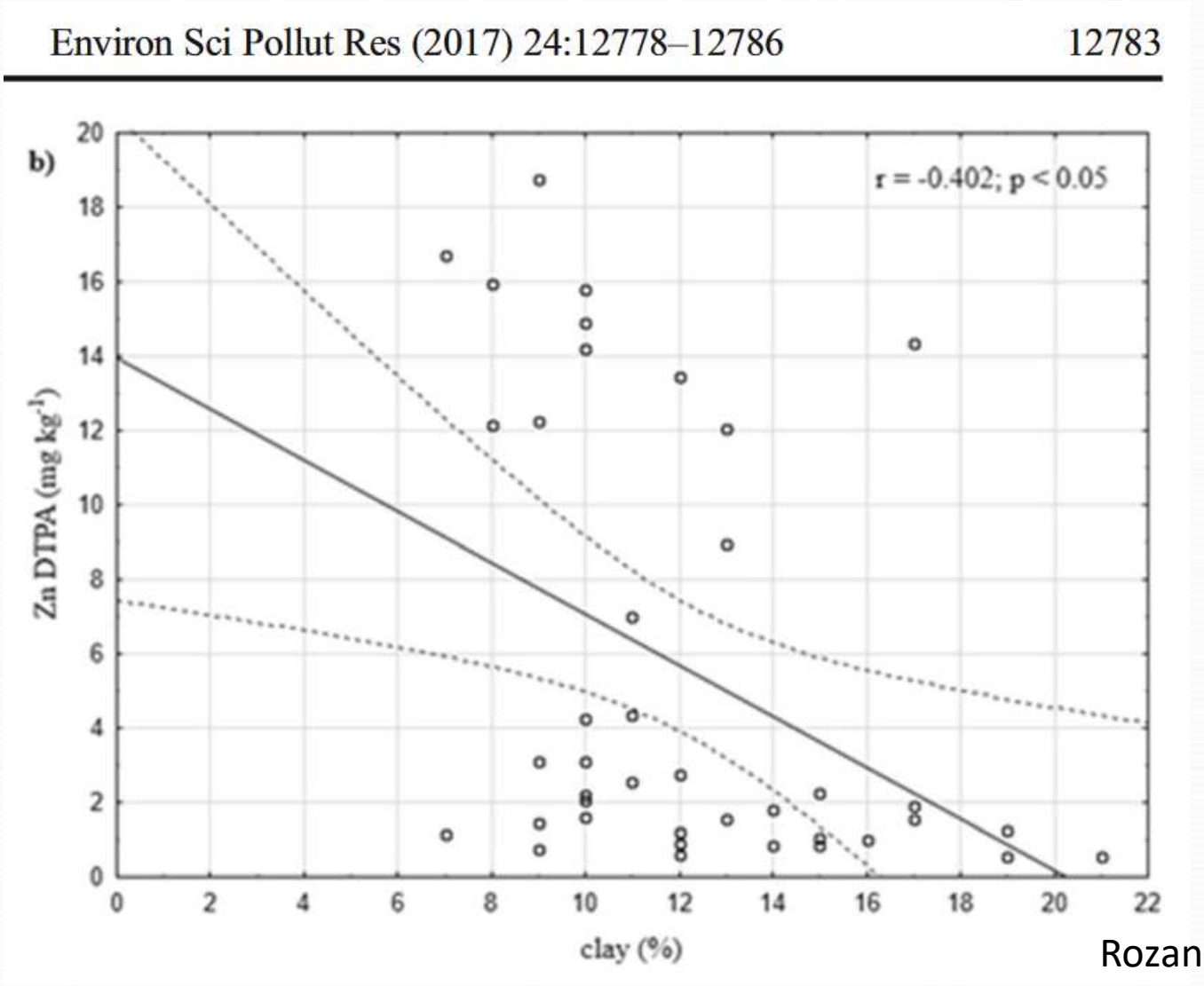
À vous : commentaires ?



Supplementary Figure 3 Correlation between actual ratings of trustworthiness and dominance in the three databases providing subjective ratings of trustworthiness and dominance (the Chicago Face Database, the Oslo Face Database and the Karolinska Face Database) and the recovered trustworthiness (**A**) and dominance (**B**) levels using the Facial Action Units detected by Open Face and our random-forest model. Source data are provided as raw data and scripts on the online depository.

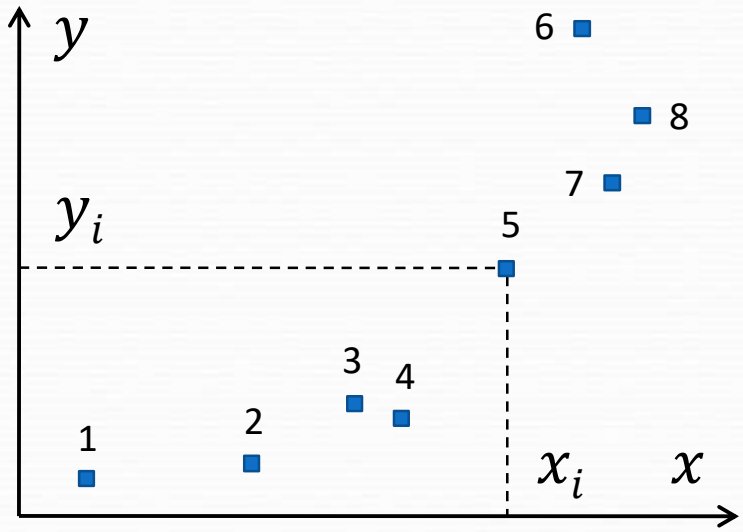
Safra et al., 2020, *Nature*

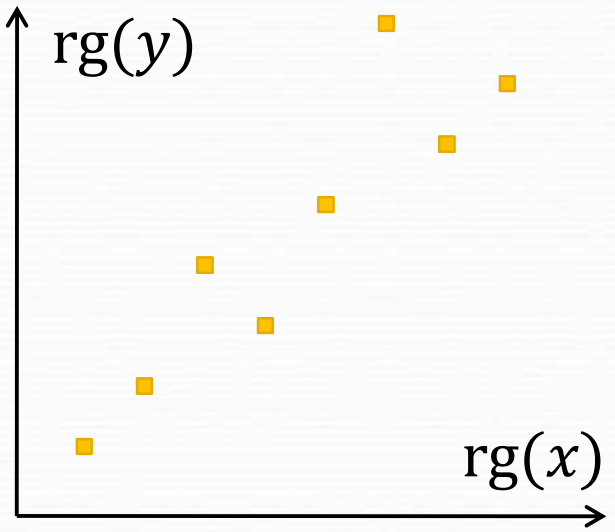
À vous : commentaires ?



Rozanski et al., 2017

Et si la relation n'est pas linéaire ?





x_1	1^{er}	y_1	1^{er}
x_2	$2^{ème}$	y_2	$2^{ème}$
x_3	$3^{ème}$	y_3	$4^{ème}$
x_4	$4^{ème}$	y_4	$3^{ème}$
x_5	$5^{ème}$	y_5	$5^{ème}$
x_6	$6^{ème}$	y_6	$8^{ème}$
x_7	$7^{ème}$	y_7	$6^{ème}$
x_8	$8^{ème}$	y_8	$7^{ème}$

Coefficient de **Spearman**

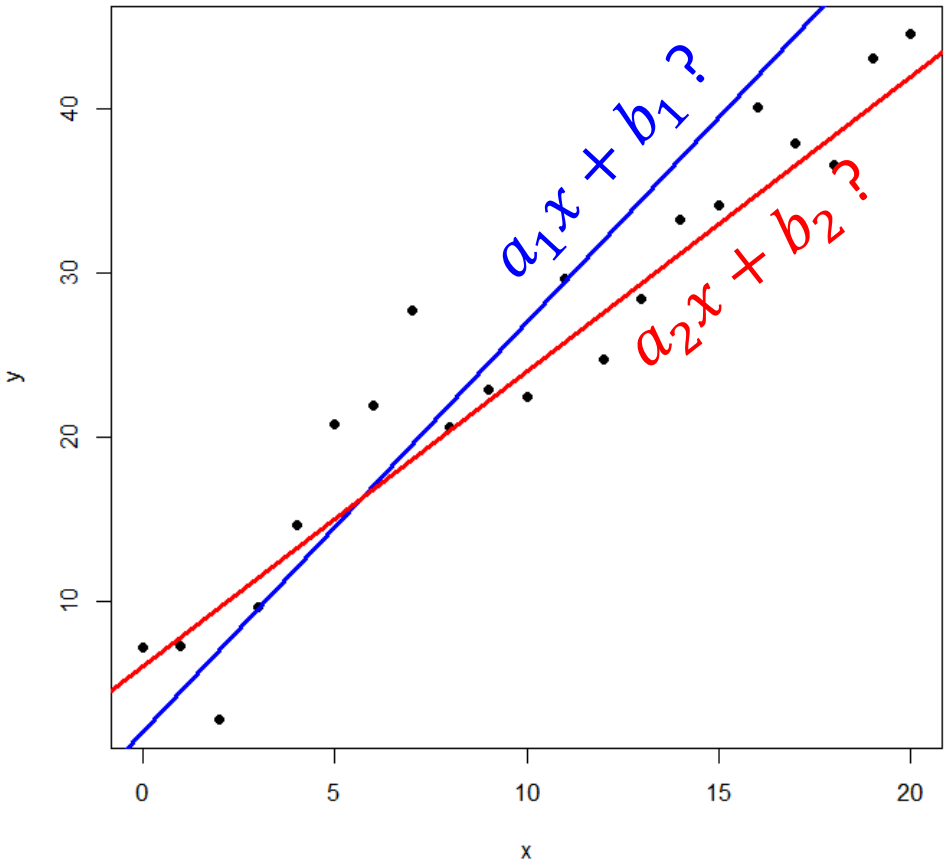
$$\rho_{XY} = r_{\mathrm{rg}(X),\mathrm{rg}(Y)}$$

$|\rho_{XY}| = 1 \Longrightarrow$ relation X–Y strictement monotone

Vers la régression linéaire

- ~~Quelle est la meilleure droite ?~~
- Quelle est la plus proche des points expérimentaux ?

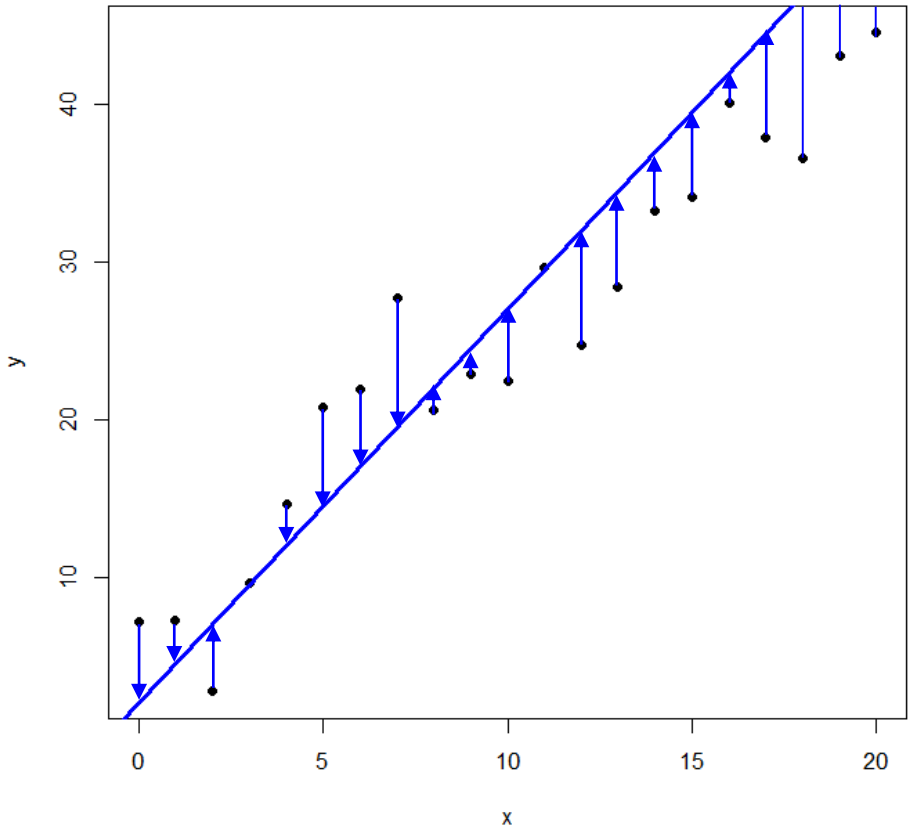
Toujours pas très clair…

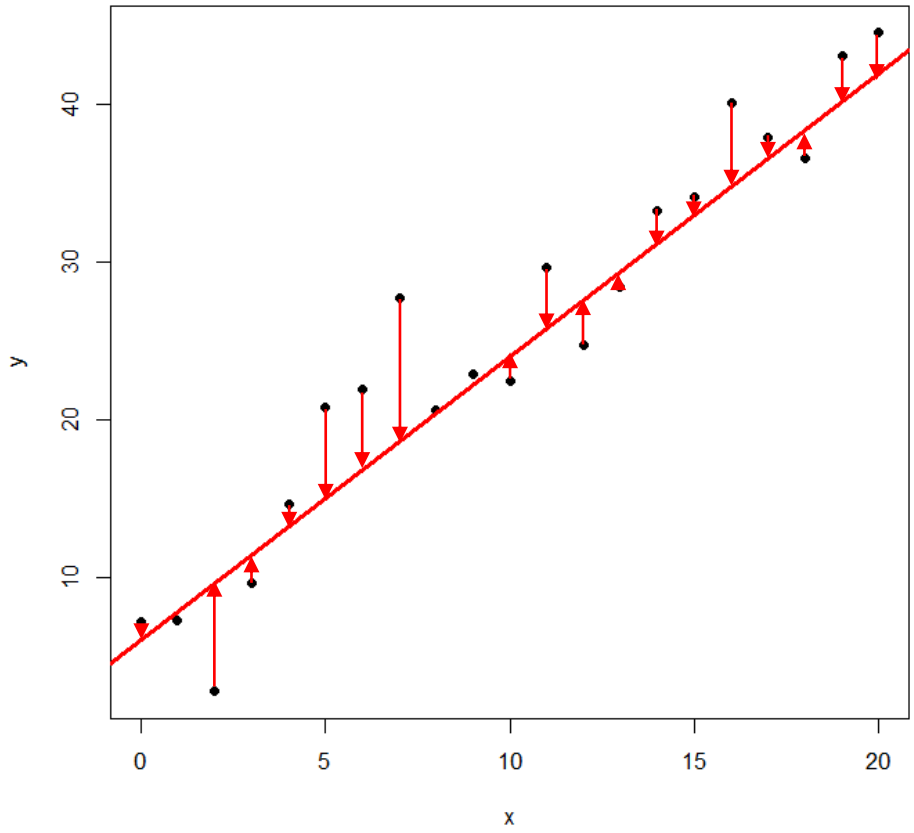


Vers la régression linéaire

$$e_i = \underbrace{(ax_i + b)}_{\text{Modèle}} - \underbrace{y_i}_{\text{Obs.}}$$

Critère **agrégé** ?





Vers la régression linéaire

- Adéquation parfaite : tous les e_i sont nuls.

$$\Leftrightarrow \quad \sum_{i=1}^{n} {e_i}^2 = 0 \quad \text{ou} \quad \sum_{i=1}^{n} |e_i| = 0 \quad \text{ou...}$$

- **Distance** mathématique entre modèle et observations
- La qualité du modèle est d'autant meilleure que cette distance est proche de zéro
- **« Moindres carrés »** $\Longrightarrow \quad \min \sum_{i=1}^{n} \big((ax_i + b) - y_i\big)^2$
- Pas la seule possibilité !

Vers la régression linéaire

- Le problème linéaire (avec critère des moindres carrés) peut se résoudre analytiquement
- La différentielle s'annule au minimum :

$$\begin{cases} \dfrac{\partial \mathrm{SCE}}{\partial a} = 0 \quad \Longrightarrow \quad \displaystyle\sum_{i=1}^{n} 2x_i(ax_i + b - y_i) = 0 \\ \dfrac{\partial \mathrm{SCE}}{\partial b} = 0 \quad \Longrightarrow \quad \displaystyle\sum_{i=1}^{n} 2(ax_i + b - y_i) = 0 \end{cases}$$

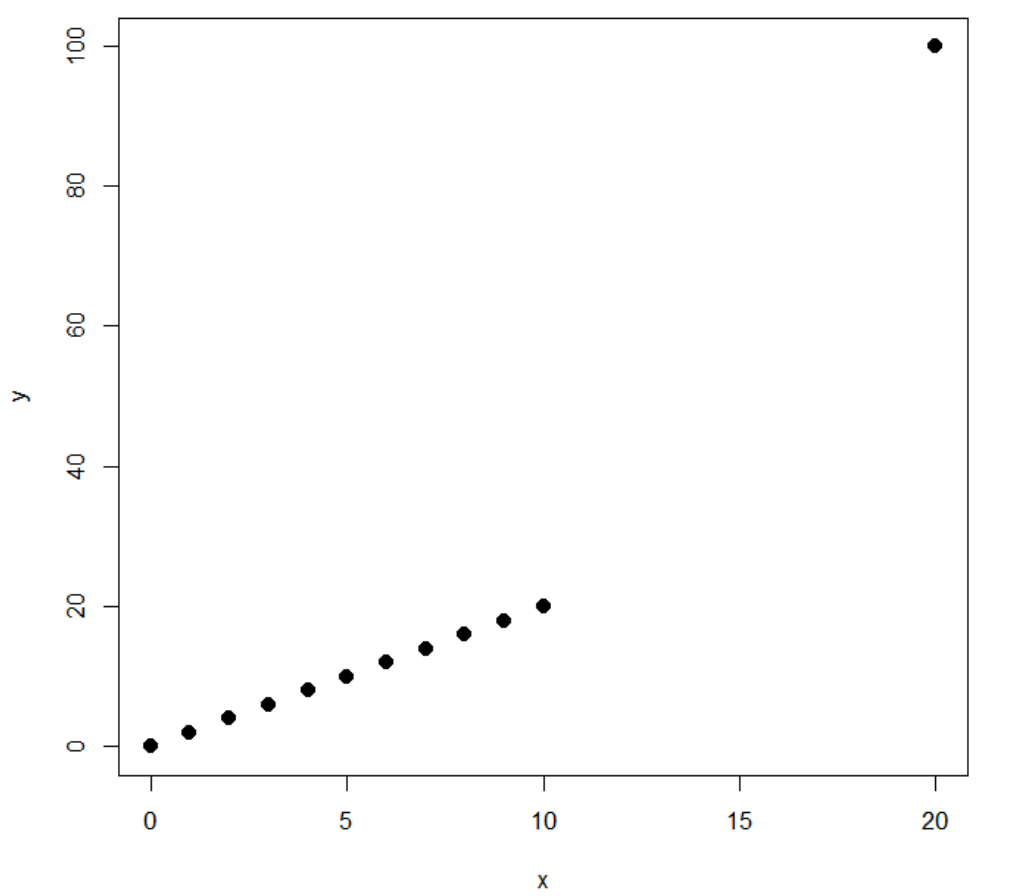
- Soit :

$$b = \bar{y} - a\bar{x} \qquad \text{et} \qquad a = \frac{\mathrm{cov}(x, y)}{\mathbb{V}(x)}$$

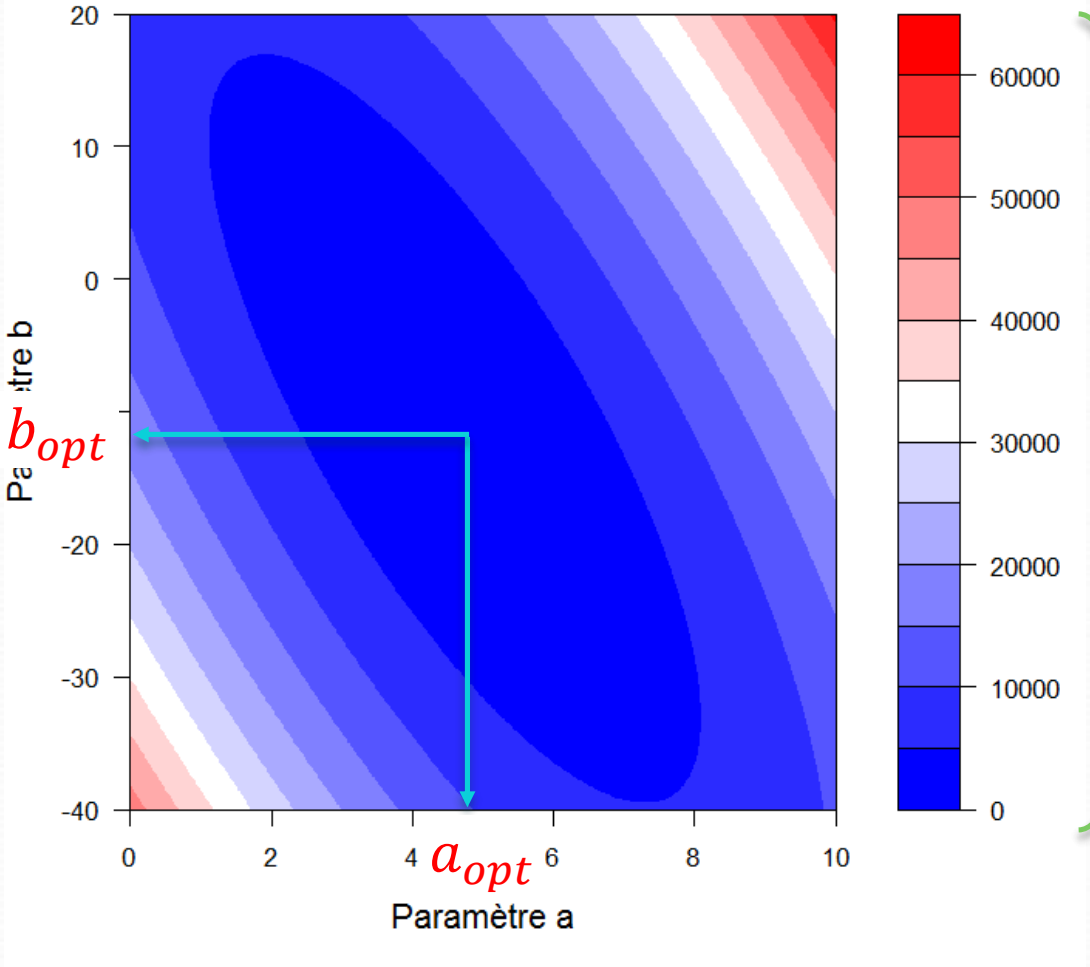
Vers la régression linéaire

Le choix du critère n'est pas anodin !

Gare aux procédures automatiques…

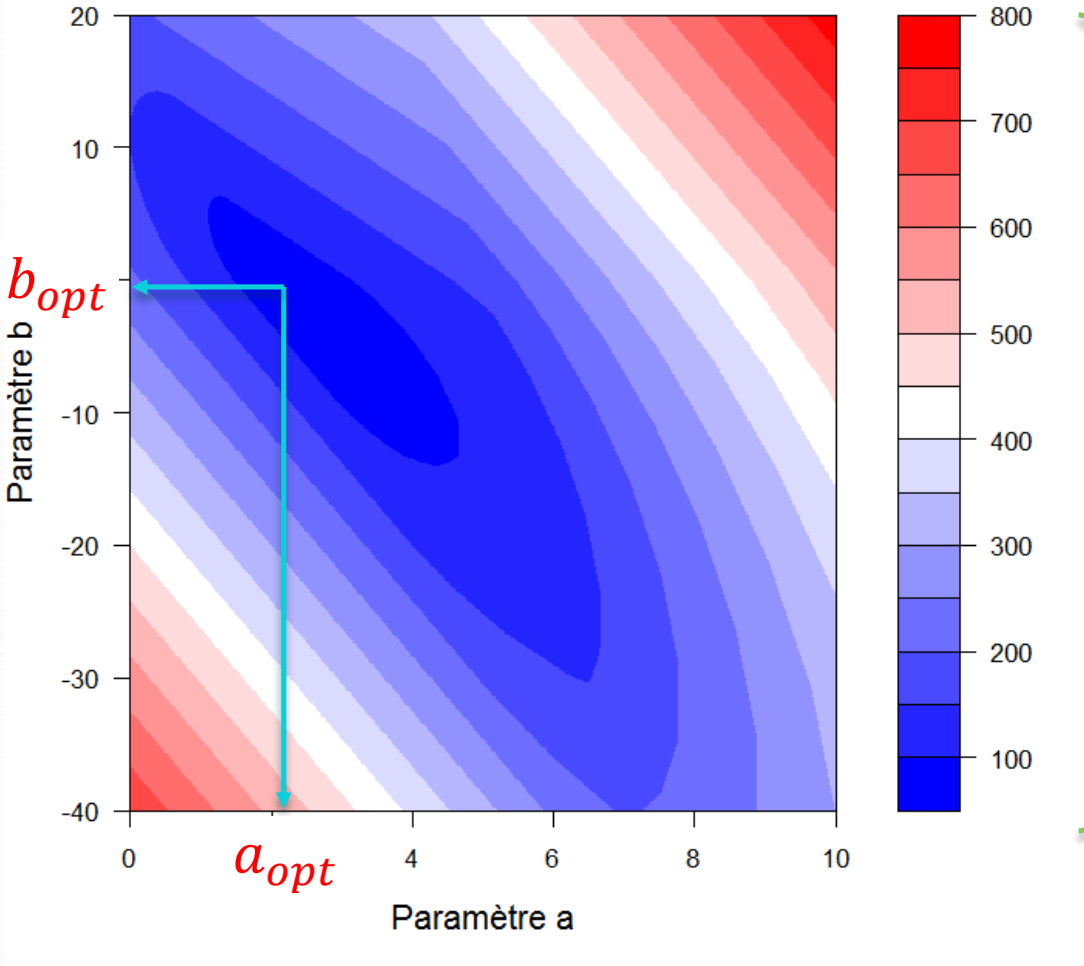


Vers la régression linéaire



$$\sum_{i=1}^{n}\left((ax_i + b) - y_i\right)^2$$

Vers la régression linéaire

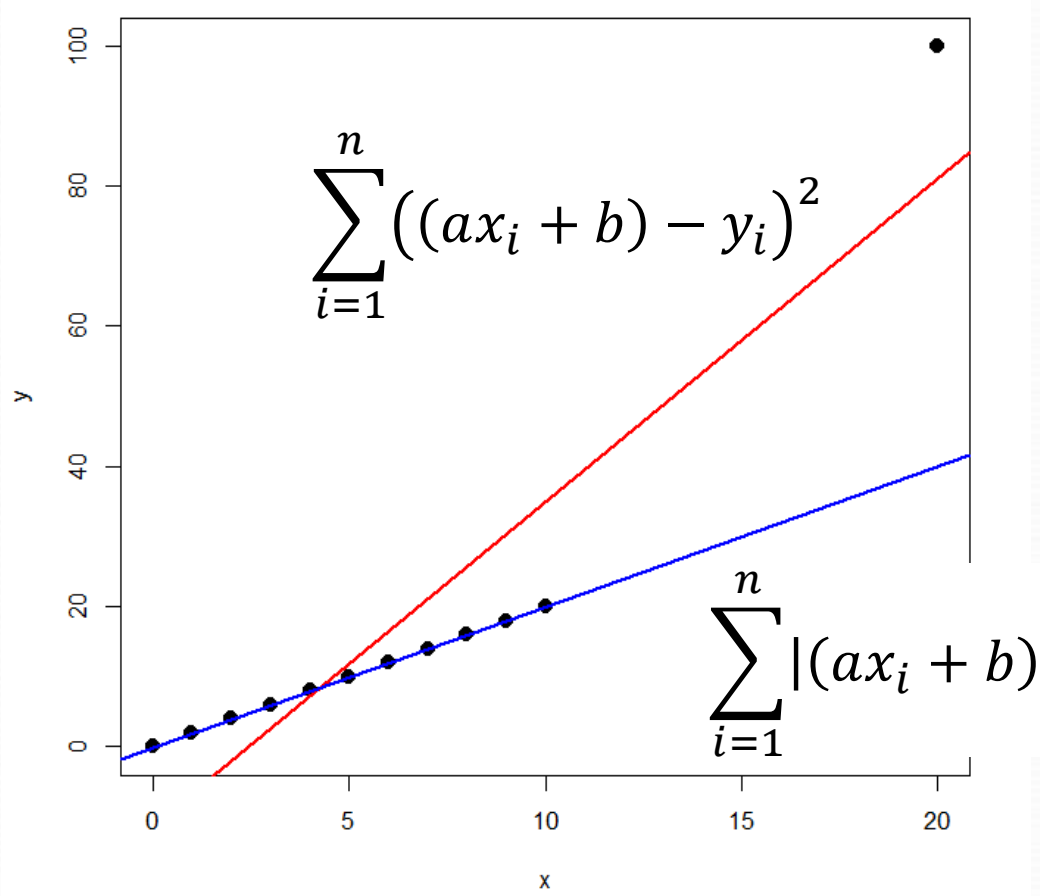


$$\sum_{i=1}^{n} |(ax_i + b) - y_i|$$

Vers la régression linéaire

Le choix du critère n'est pas anodin !

Gare aux procédures automatiques…



À vous : commentaires ?

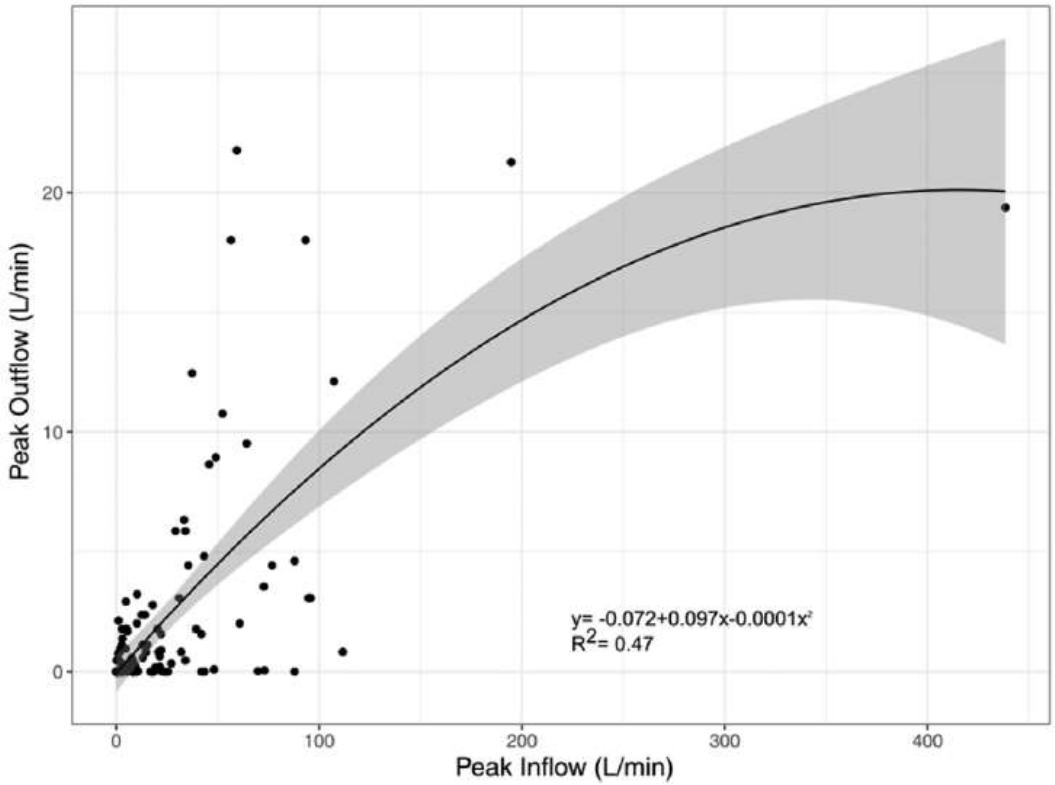
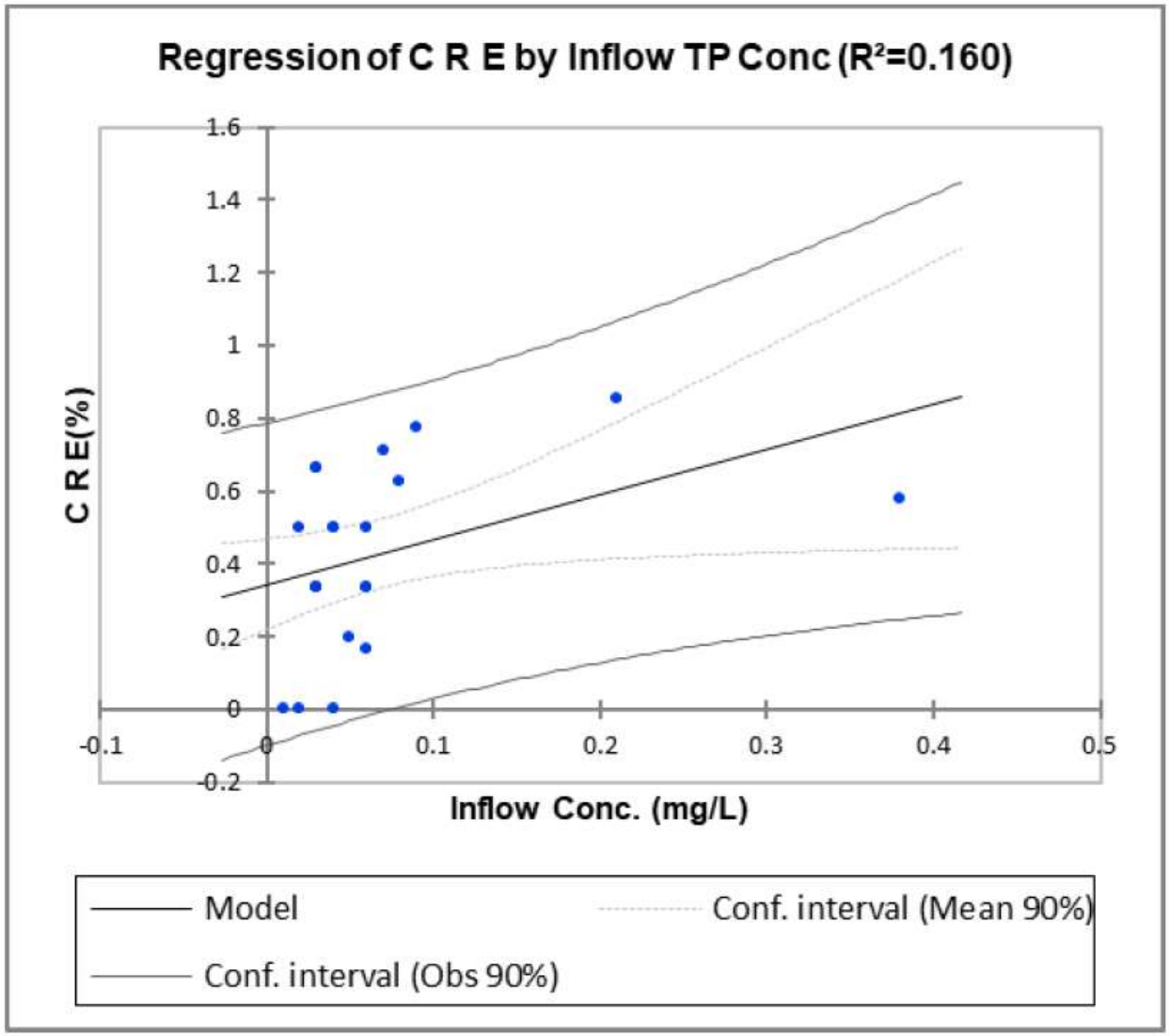


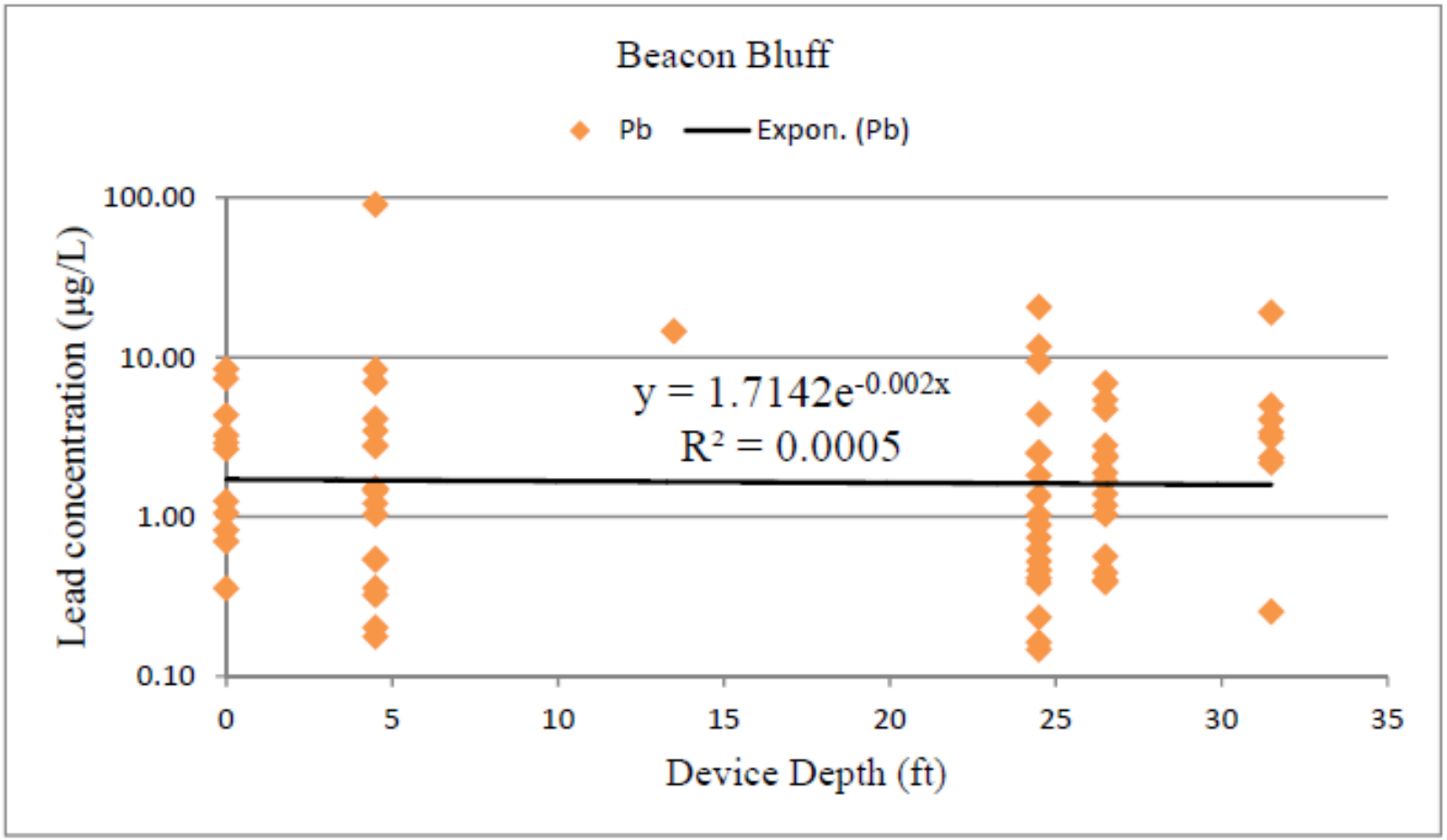
Fig. 5. Relationship between peak inflow and peak outflow rate ($L\,min^{-1}$) for the storm events sampled spanning May to October/November 2015 and 2016 in Burlington, Vermont.

À vous : commentaires ?



Drapper & Hornbuckle, 2018

À vous : commentaires ?

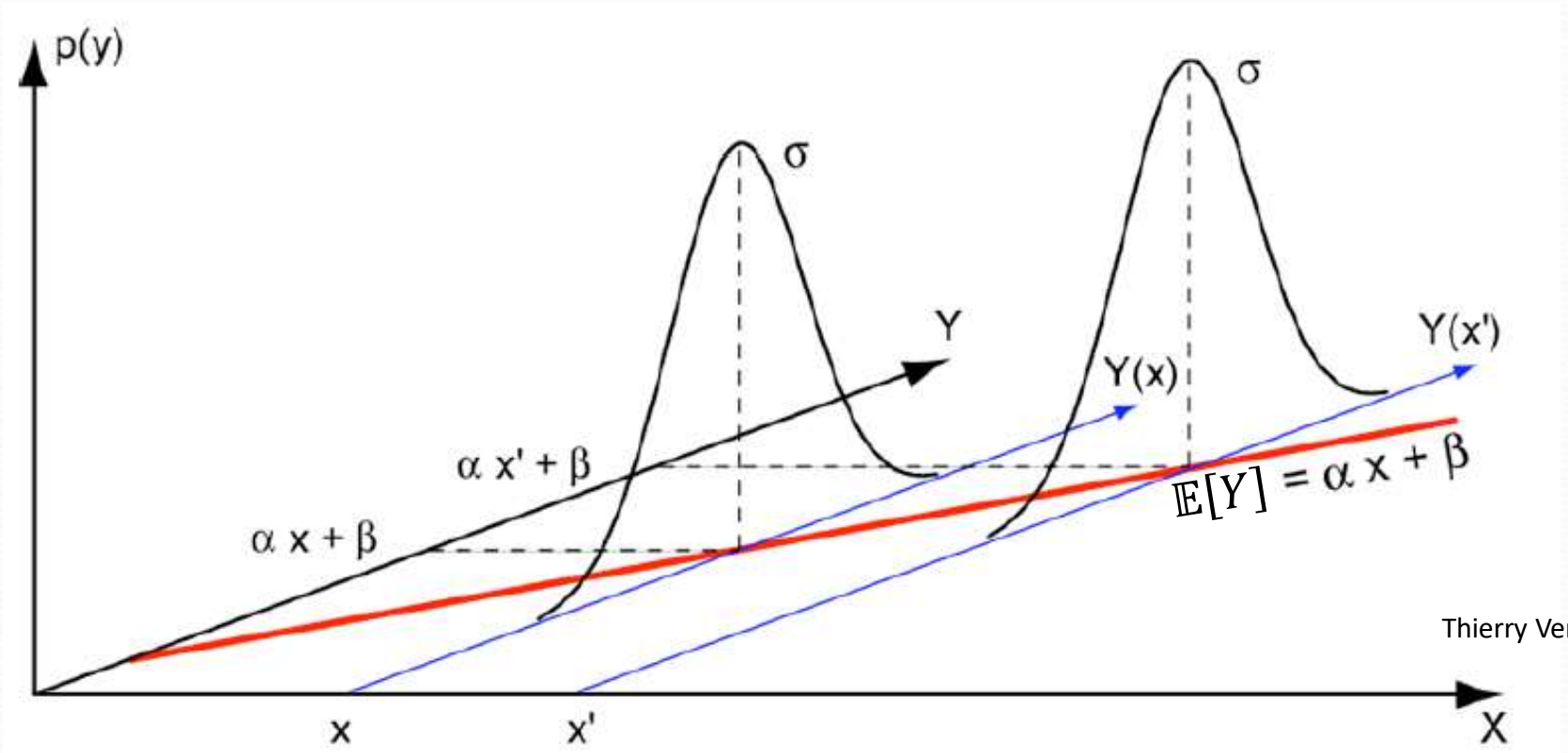


Nieber et al., 2014, *US Environmental Protection Agency*

Modèle probabiliste de la régression

$$Y_i = \alpha x_i + \beta + \varepsilon_i$$

- Y_i et ε_i sont des variables aléatoires
- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, indépendants entre eux



Thierry Verdel, 2007

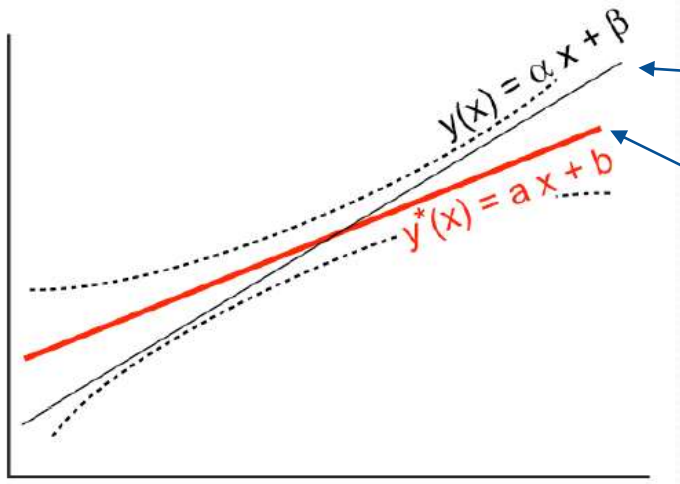
Modèle probabiliste de la régression

- La droite de régression caractérise le comportement moyen, c'est-à-dire $\mathbb{E}[Y] = \alpha x + \beta$
- Les écarts entre les points et la droite sont dus à des fluctuations aléatoires du signal
- Deux grands avantages de cette vision probabiliste :
  1. Les **estimations** des coefficients de la droite α et β sont livrées en kit avec :
     - un intervalle de confiance
     - un test de nullité
  2. La **prévision statistique** par le modèle linéaire s'assortit d'une estimation de l'incertitude

Modèle probabiliste de la régression

Attention !! Il existe deux questions distinctes…

1. La droite des moindres carrés **n'est pas** la droite de régression ($\mathbb{E}[Y] = \alpha x + \beta$), c'en est une estimation. Intervalle de confiance de la moyenne $\mathbb{E}[Y]$?



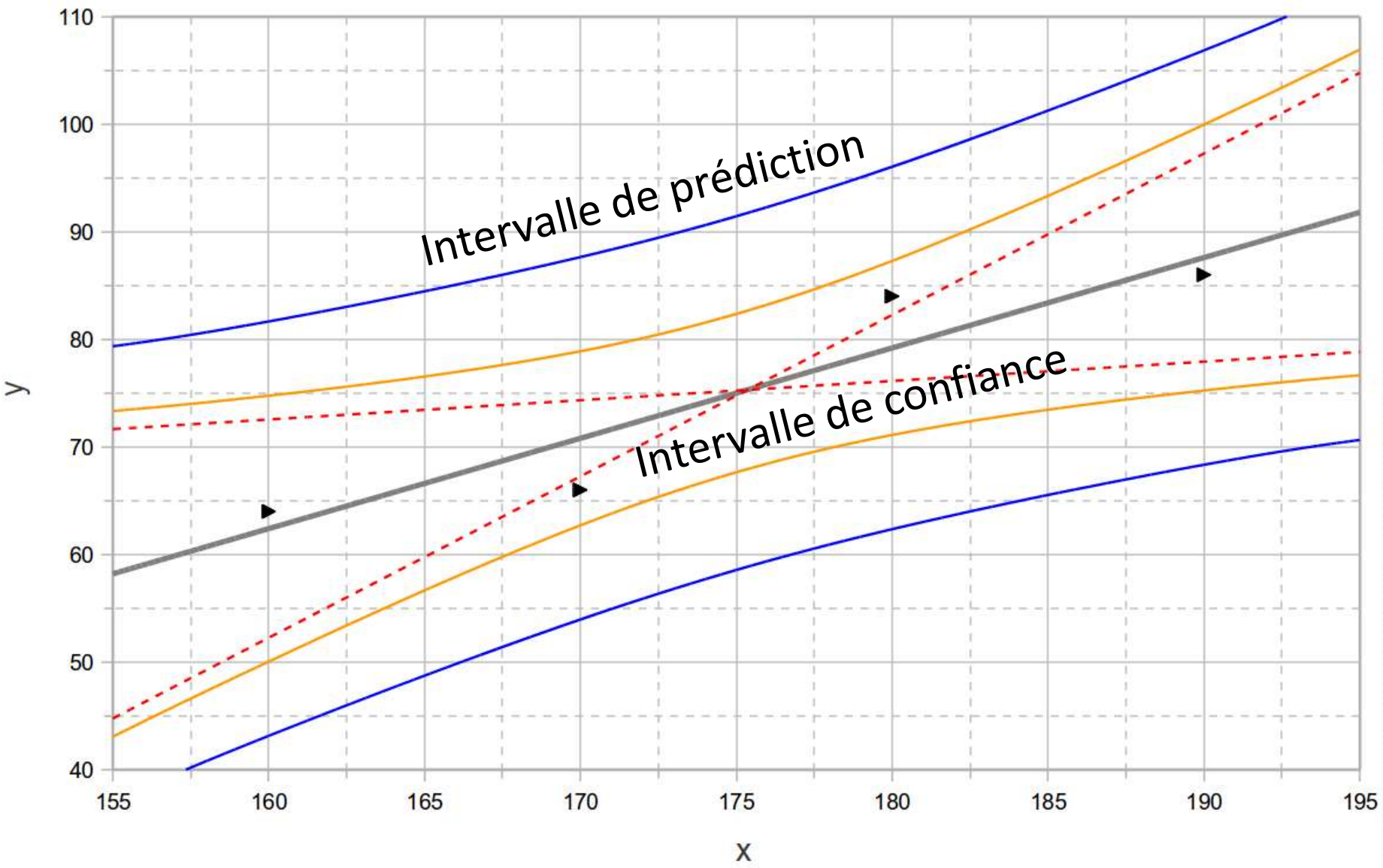
« Vraie » droite de régression $\mathbb{E}[Y]$ (inconnue)

Droite des moindres carrés

Thierry Verdel, 2007

2. On **prédit** la valeur de l'observation $y(x_0)$ par $\hat{y}(x_0) = ax_0 + b$. Intervalle de prévision ?

Modèle probabiliste de la régression



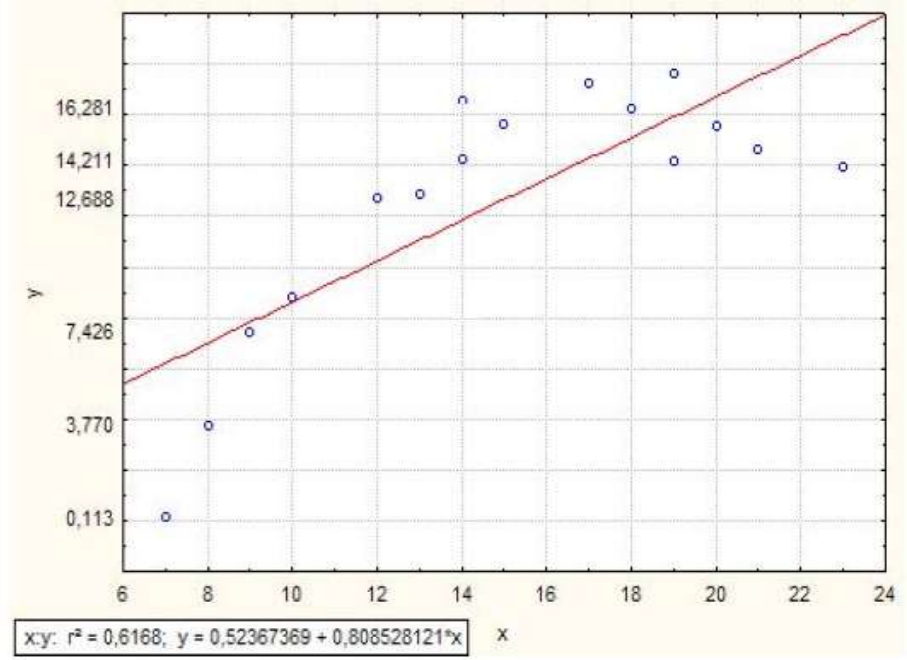
Quelques mises en garde

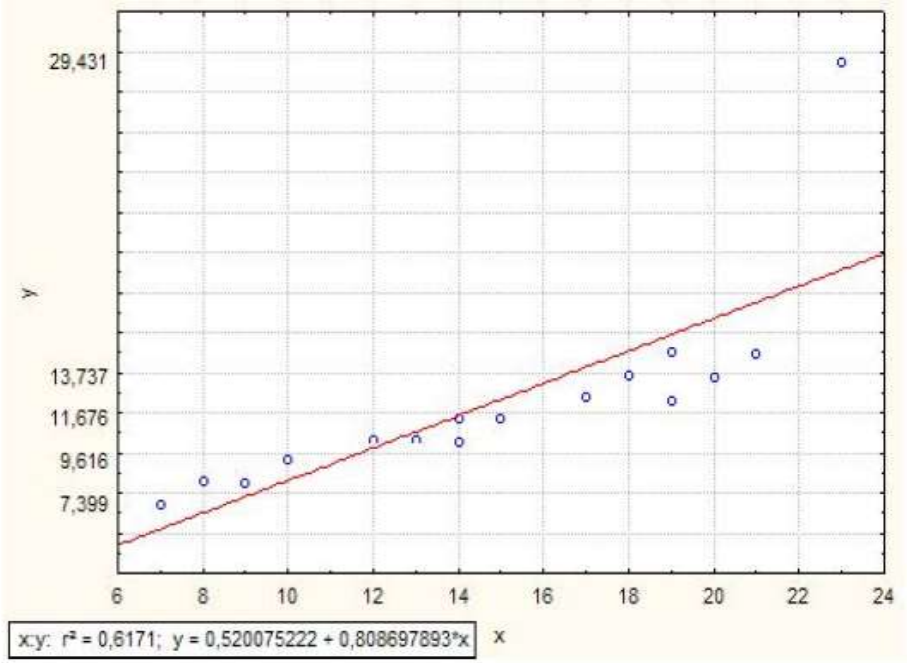
- Par rapport à la vision déterministe, on gagne de l'information en émettant des hypothèses statistiques :

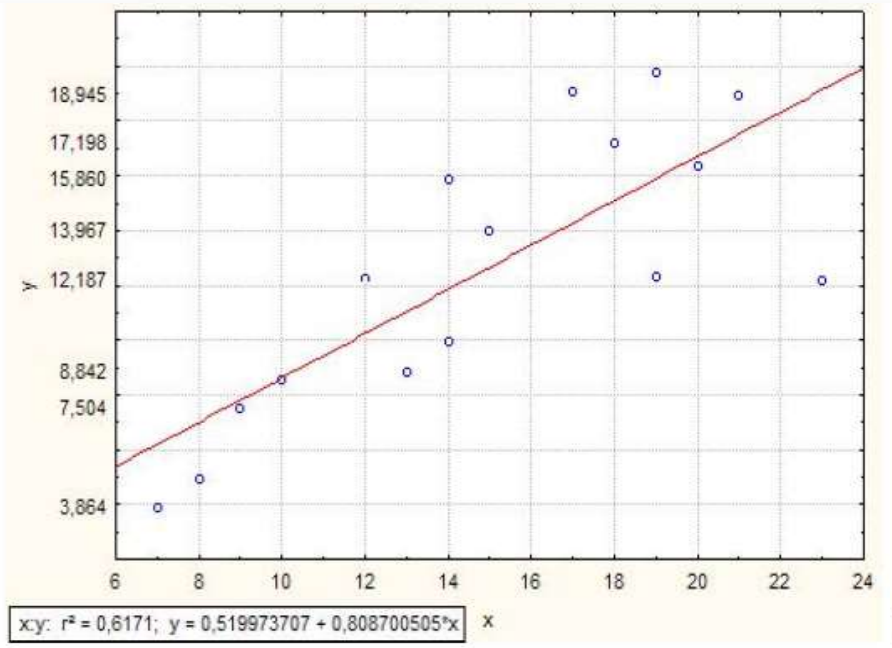
$$Y_i = \alpha x_i + \beta + \varepsilon_i \text{ avec}$$

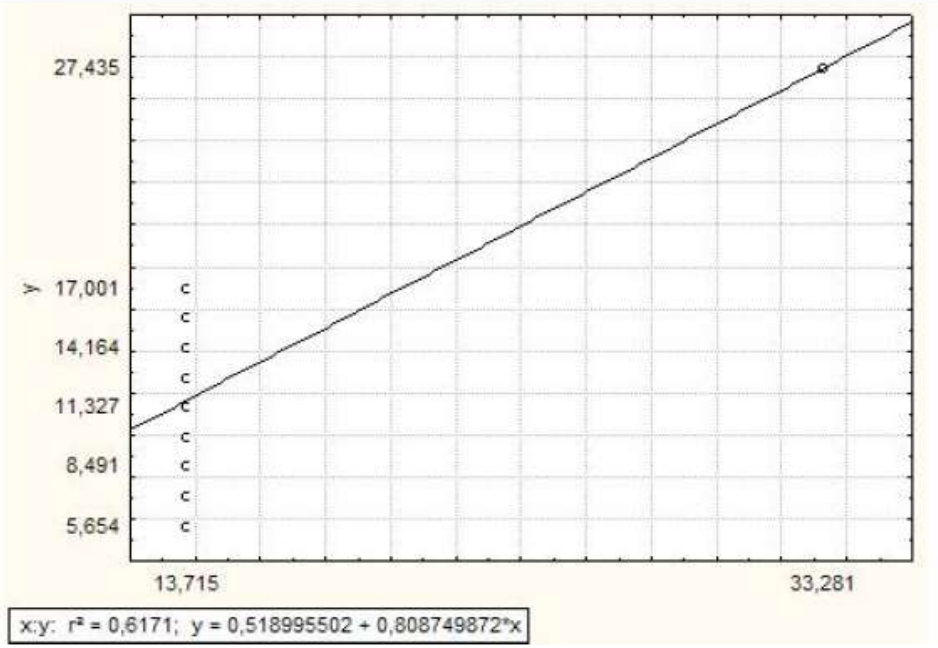
  - $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$: même variance pour tous les résidus (« homoscédasticité »)
  - indépendants entre eux : pas d'autocorrélation
  - et indépendants de x (sinon plus de modèle linéaire…)

- Le**diagnostic des résidus** est une étape fondamentale !
- Sinon toutes les décisions issues des tests et tous les intervalles de confiance n'ont plus de légitimité…









Tomassone et al., 1983

À retenir

- Approche « initiale » de la régression linéaire : une démarche de calage
- Qualité de l'ajustement : $R^2 =$ fraction de la variance totale expliquée par le modèle linéaire
- Si on commence à faire des probas :
  - Tests d'hypothèses sur les coefficients a et b
  - Intervalle de confiance de la droite de régression
  - Intervalle de prévision d'une valeur $\hat{y}(x_0)$
- Attention aux extrapolations en-dehors de la gamme des mesures $x_1, \ldots, x_n$!