Séance 5. Statistique bivariée

RELATION ENTRE DEUX VARIABLES QUANTITATIVES

Organisation de la séance

1 – L’intérêt de la statistique bivariée en sciences sociales

2 – La distinction entre corrélation et causalité

3 – La représentation graphique de deux variables quantitatives : le nuage de points

4 – Droite de régression et linéarité

5 – La mesure de l’intensité d’une relation linéaire : le coefficient de corrélation (r de Pearson)

6 – Exercices d’application

La statistique bivariée en SHS

Sciences sociales cherchent à **comprendre et expliquer des phénomènes sociaux complexes** à partir des **liaisons multiples** qui existent entre :

- Acteurs et structures sociales; individus et groupes sociaux; comportements et relations sociales; perceptions et actions; actions et résultats; etc.
- Exemples : Origine sociale et réussite scolaire; âge et participation politique; comportement de consommation et lieu de résidence; statut socioéconomique et santé; diplôme et revenu; investissement et profit, etc.

L'étude des liaisons entre variables permet de **tester les théories et les hypothèses** proposées dans les sciences sociales (**logiques sociales entre les variables**).

- Validation empirique (preuve statistique) renforce **la crédibilité des connaissances** dans le domaine.
- Validation empirique aide à **réduire les biais et les préjugés** dans la compréhension des phénomènes sociaux (enjeu de la distinction corrélation-causalité).

Sciences sociales cherchent à étudier les « variations concomitantes » (Durkheim dans les Règles de la méthode sociologique) entre situations, phénomènes (les relations, dépendances, corrélations entre variables)

La distinction corrélation/causalité

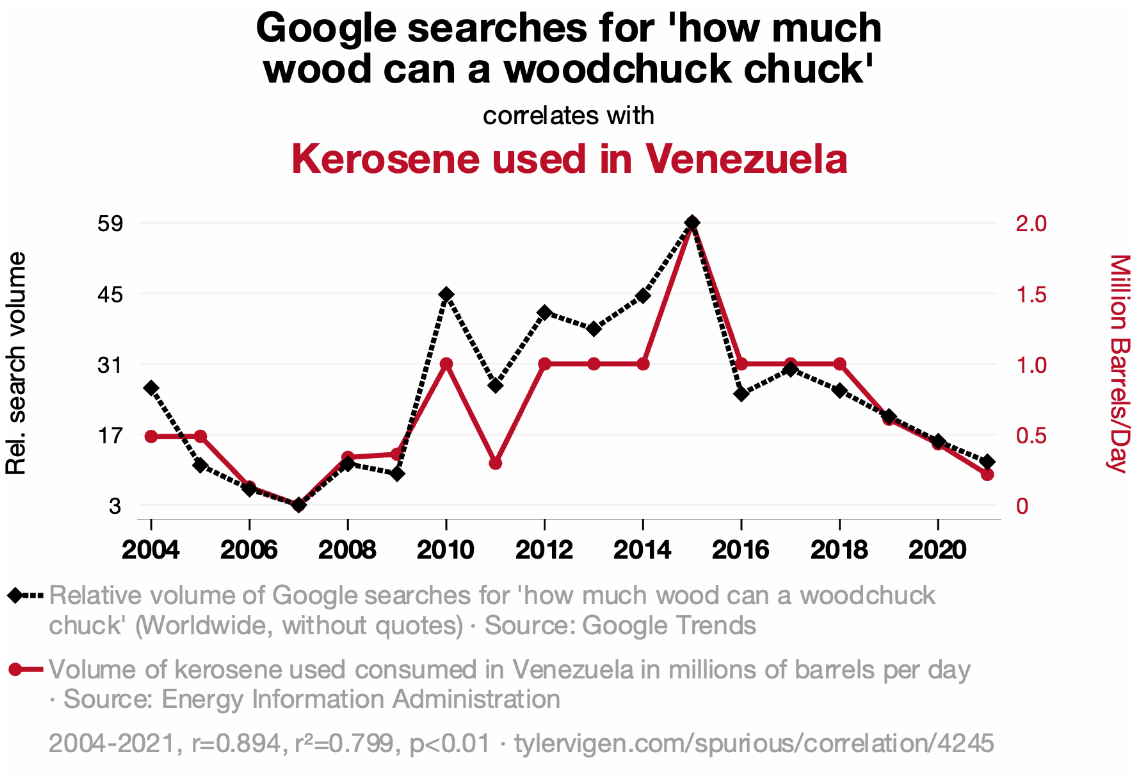
Corrélation *(terme souvent utilisé de manière approximative ou abusive dans le langage courant).*

- **Renvoie au concept de liaison :** *relier, mettre en correspondance deux choses ou deux évènements*
  - Les liaisons peuvent être de natures différentes : physiques (lien taille/poids); logiques (cas des SHS - sous certaines hypothèses théoriques, liens entre revenu et consommation par exemple)
  - Les liaisons sont rarement parfaites (il existe des riches avares !)
  - Les liaisons peuvent être « dangereuses »: un petit tour sur le site « spurious correlation » !

Enjeu statistique : identifier et mesurer la « force » de ces liaisons

La distinction corrélation/causalité

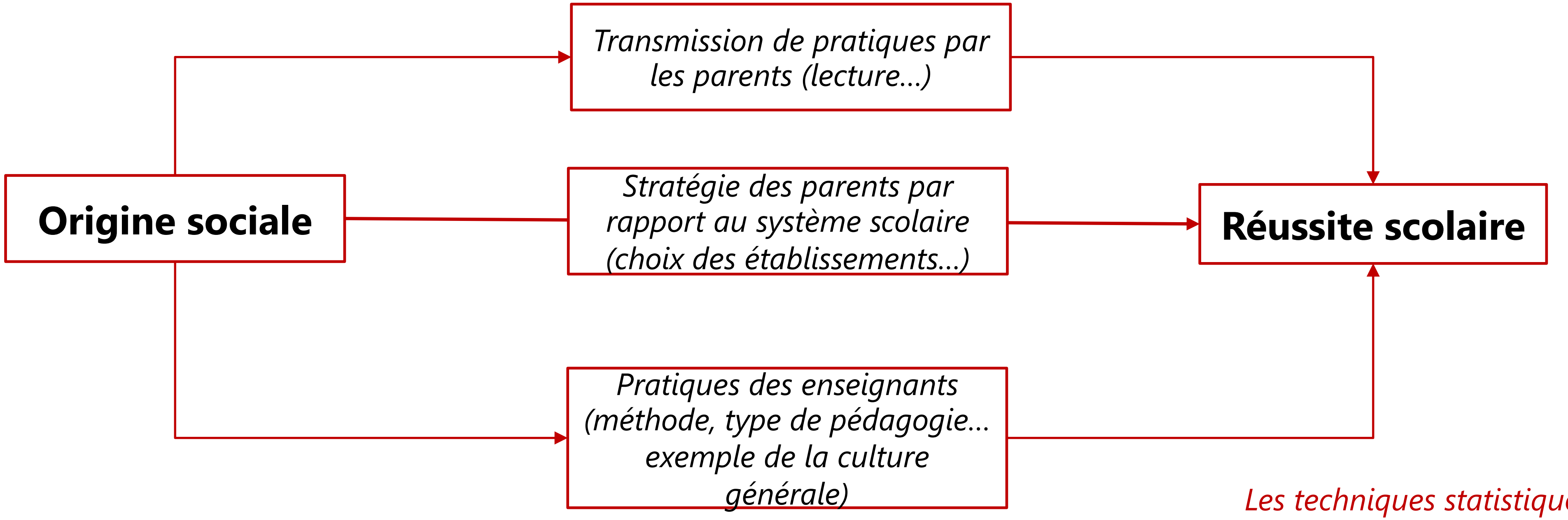
Un exemple de « spurious correlation »



La distinction corrélation/causalité

Une corrélation pour des causes multiples ! Un exemple

Les processus à l'œuvre derrière une corrélation entre deux phénomènes sociaux sont souvent multiples et ne se réduisent pas à une cause simple !



Les techniques statistiques doivent être multiples et combinées avec des hypothèses et savoirs (plus ou moins stabilisés) pour rendre compte d'un phénomène social !

Source : adapté de [Grossetti (2023)]

La distinction corrélation/causalité

De la corrélation à la causalité

- Etude de la corrélation n’apporte pas beaucoup de réponses sur la causalité des phénomènes
  - En particulier dans les études en coupe transversale; les données longitudinales permettent de dresser des interprétations plus robustes en matière de causalité
- L’ambition première de l’étude des corrélations bivariées n’est pas d’expliquer la causalité…
  - … mais d’identifier les liaisons entre deux phénomènes et d’en mesurer l’intensité, la force
  - … ce qui reste évidemment une première étape indispensable pour l’analyse plus fine des mécanismes causaux par la suite

Le nuage de points (diagramme de dispersion)

Représentation graphique de la relation entre deux variables x et y

- y = variable à expliquer, ou **variable dépendante**
- x = variable explicative, ou **variable indépendante**

Ces deux variables sont représentées sur un plan à deux dimensions :

- Un axe horizontal, appelé **axe de abscisses** ou axe des x
- Un axe vertical, appelé **axe des ordonnées**, ou axe des y

Un point du nuage = un **individu <u>statistique</u>** (l'unité dépend de la base de données : une personne, une entreprise, un pays, etc.)

Chaque point est positionné selon ses **coordonnées** $(x_i; y_i)$

Taux de mortalité infantile et PIB/hab

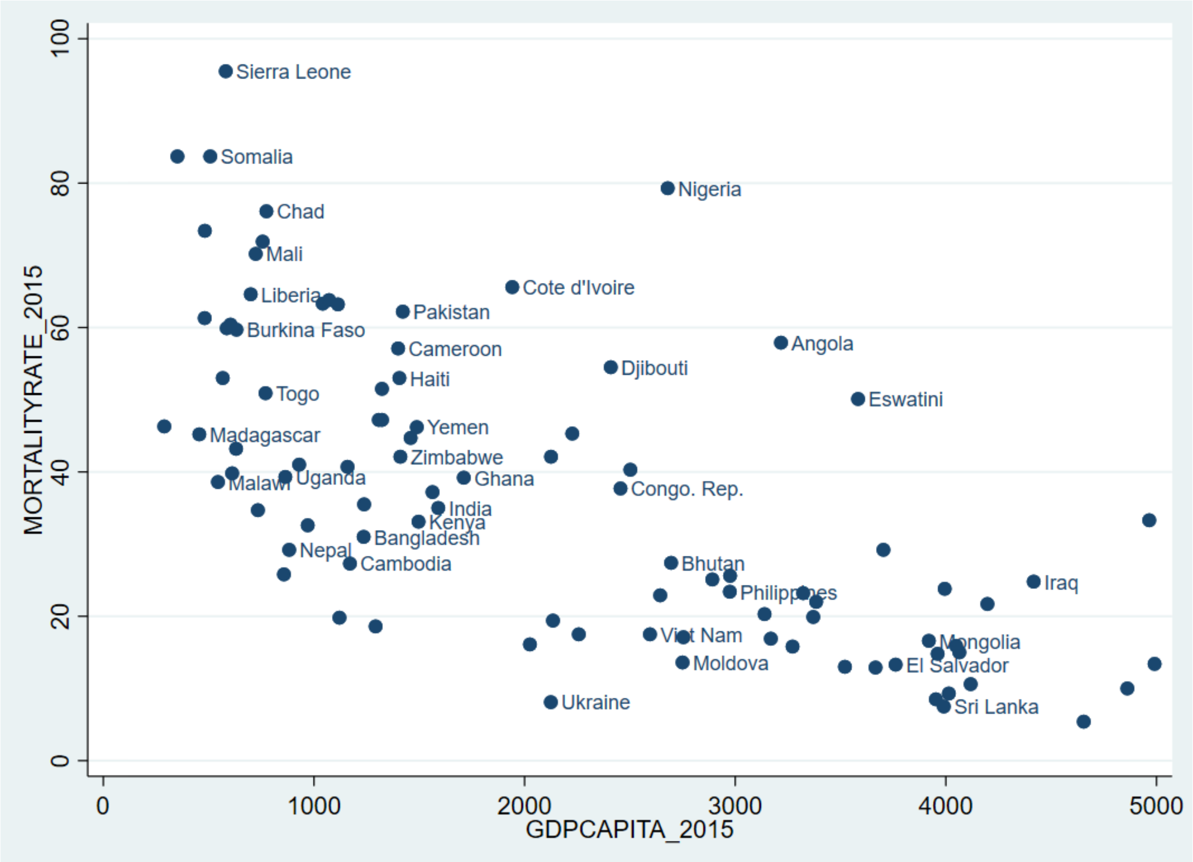
Données de 2015 pour tous les pays dont le PIB/hab < 5000$ (N=93) – Source : Banque Mondiale, WDI Databank

Deux variables :

- **Axe des abscisses (X)** : PIB/hab en dollars
- **Axe des ordonnées (Y)** : Taux de mortalité infantile (nb de décès d'enfants de moins d'un an, ‰)

Lecture du nuage de points :

- Quel pays avait le taux de mortalité infantile le plus élevé en 2015?
- Quelles sont les coordonnées du Sri Lanka?
- Comment décririez-vous la relation entre les deux variables?



La droite de régression

Une tendance générale peut être tracée grâce à la **droite de régression linéaire** ($y = ax + b$)

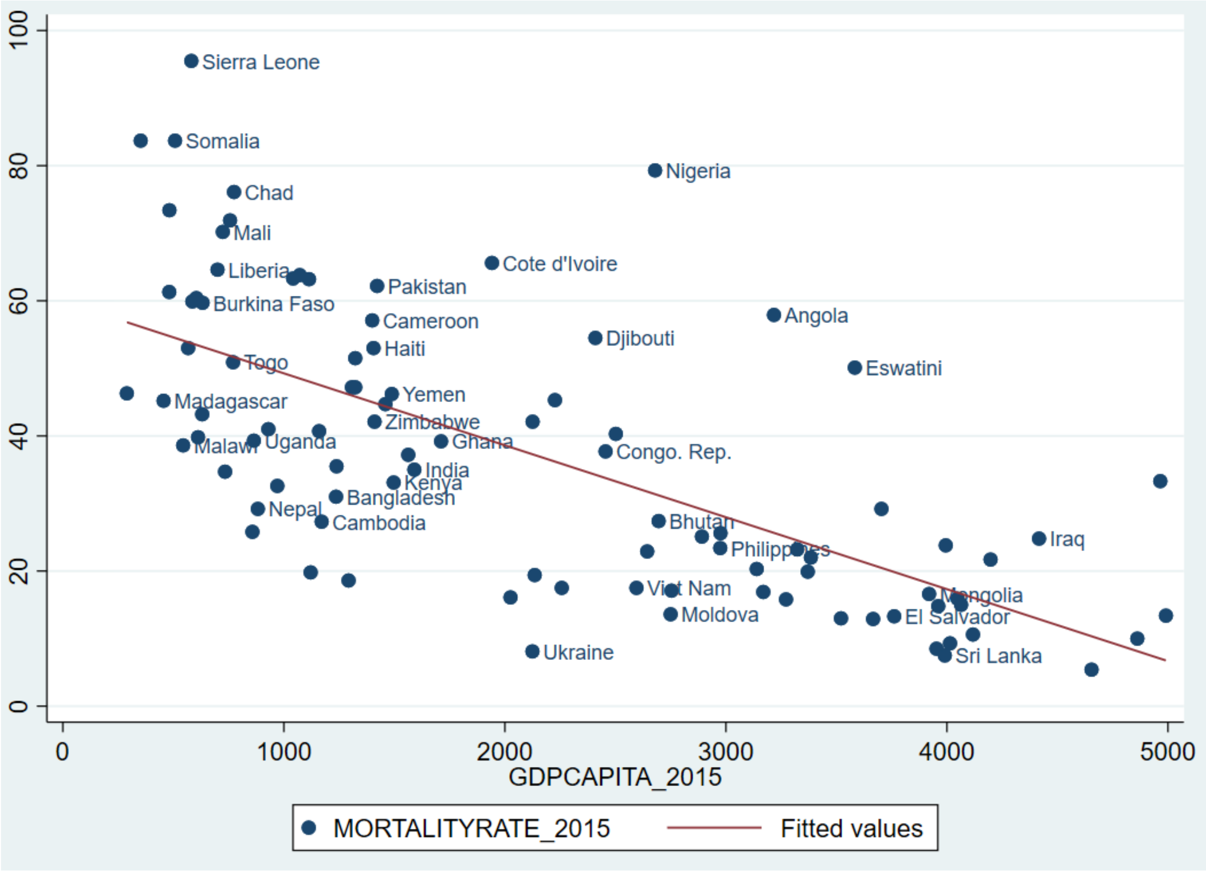
Elle est utile pour faire des **prédictions** : si un pays a un niveau x_i (<u>valeur réelle</u>) de PIB/hab alors la droite nous donne une estimation de son taux de mortalité infantile $\hat{y}_i$ (<u>valeur prédite</u>) – On ne peut pas parler de **causalité!**

La différence entre les valeurs prédites et réelles est appelée **résidu** ($\varepsilon_i = y_i - \hat{y}_i$)).

Plus les résidus sont importants – ou la **dispersion** des points est élevée – moins la relation (x;y) est forte

<u>Lecture du nuage de points :</u>

- Quelle est la valeur prédite du taux de mortalité infantile ($\hat{y}_i$) pour Djibouti? Quelle est sa valeur réelle (y_i)?
- Que peut-on en déduire concernant le résidu?



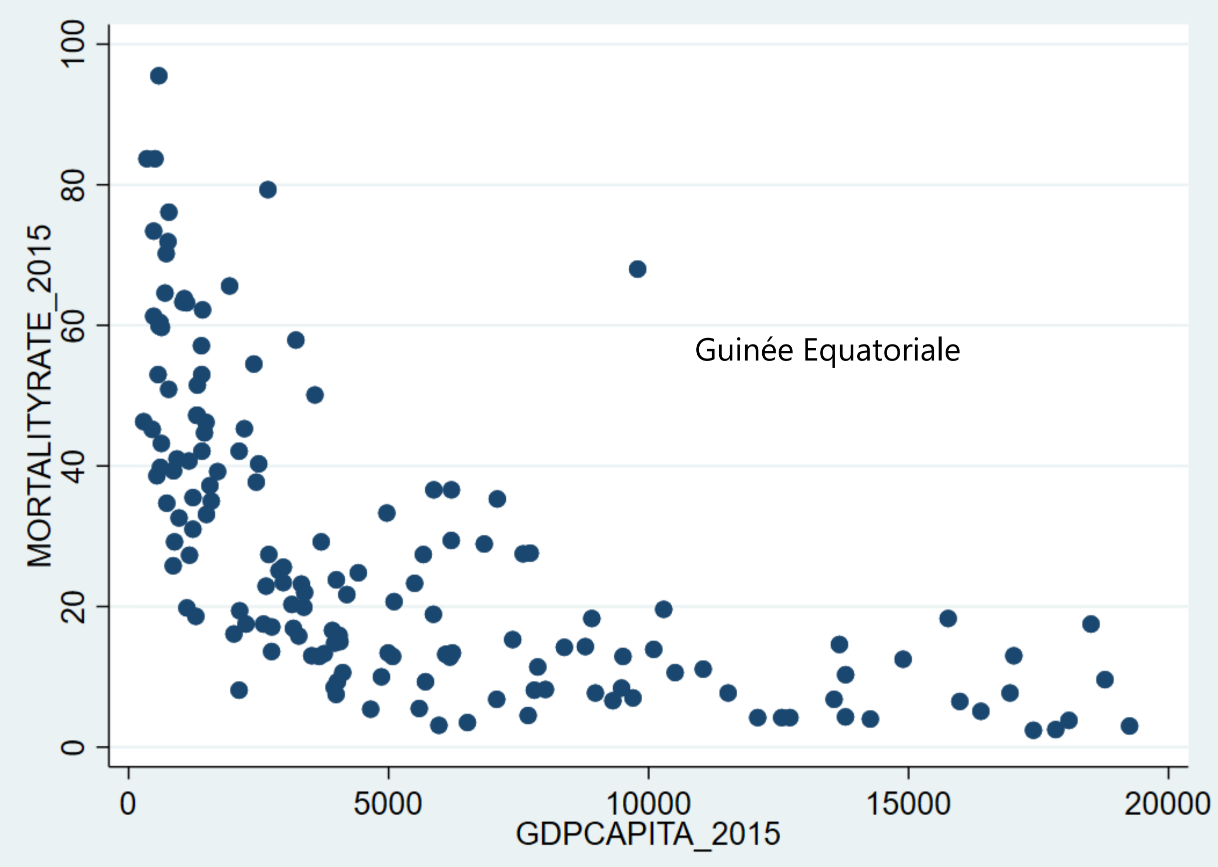
Linéarité *vs* non linéarité

La prédiction est une **approximation** : le PIB/hab seul n'est pas suffisant pour prédire avec précision le taux de mortalité infantile qui prévaut dans un pays i

Le nuage de point à droite inclus désormais en plus les pays dont le PIB/hab, en 2015, était compris entre 5000 et 20 000$ (N=157)

Que constate-t-on sur ce nouveau nuage de points?

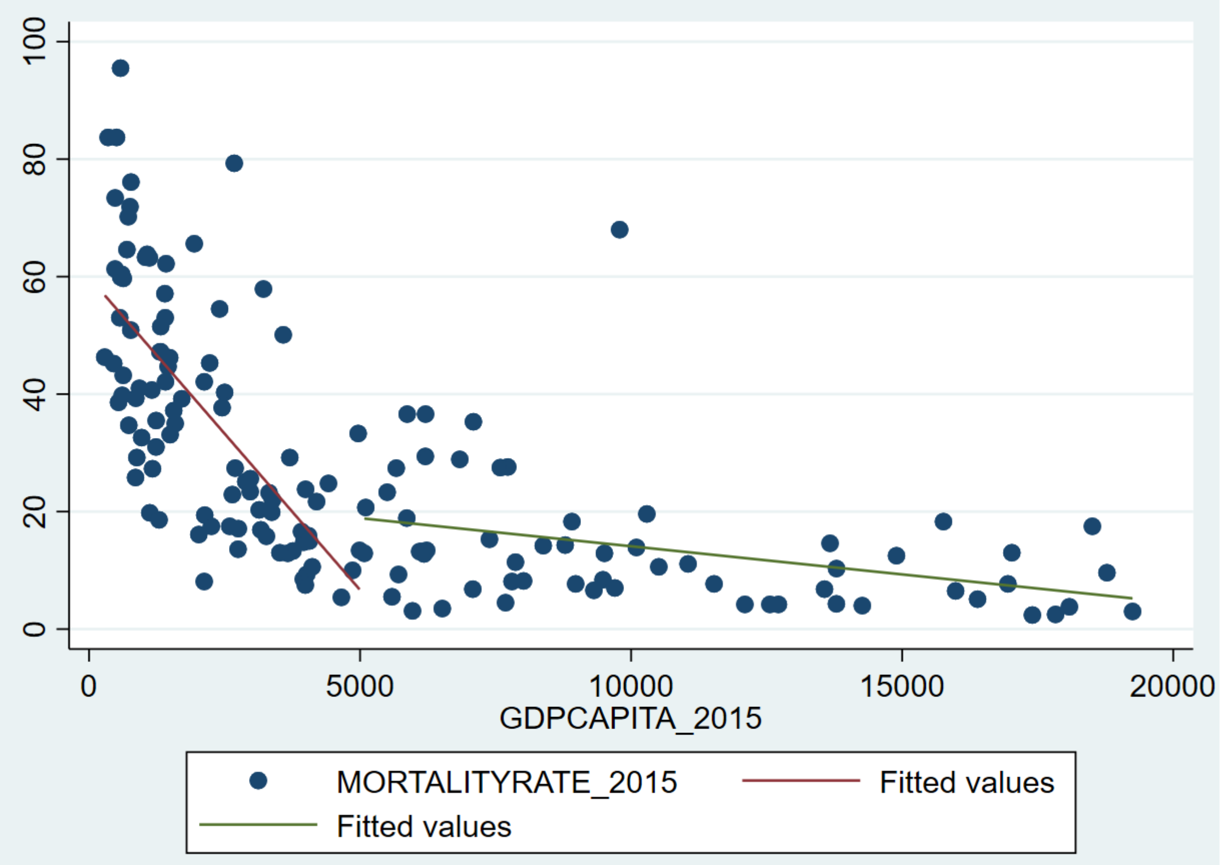
- (1) Un point semble significativement éloigné du reste de la distribution (la Guinée Equatoriale)
- On appelle ces points des **outliers** : ils peuvent exercer une influence importante sur la pente de la droite, ou sur d'autres métriques
- (2) La relation est beaucoup moins **linéaire**, même si elle semble rester **monotone** (toujours décroissante)



Linéarité *vs* non linéarité (2)

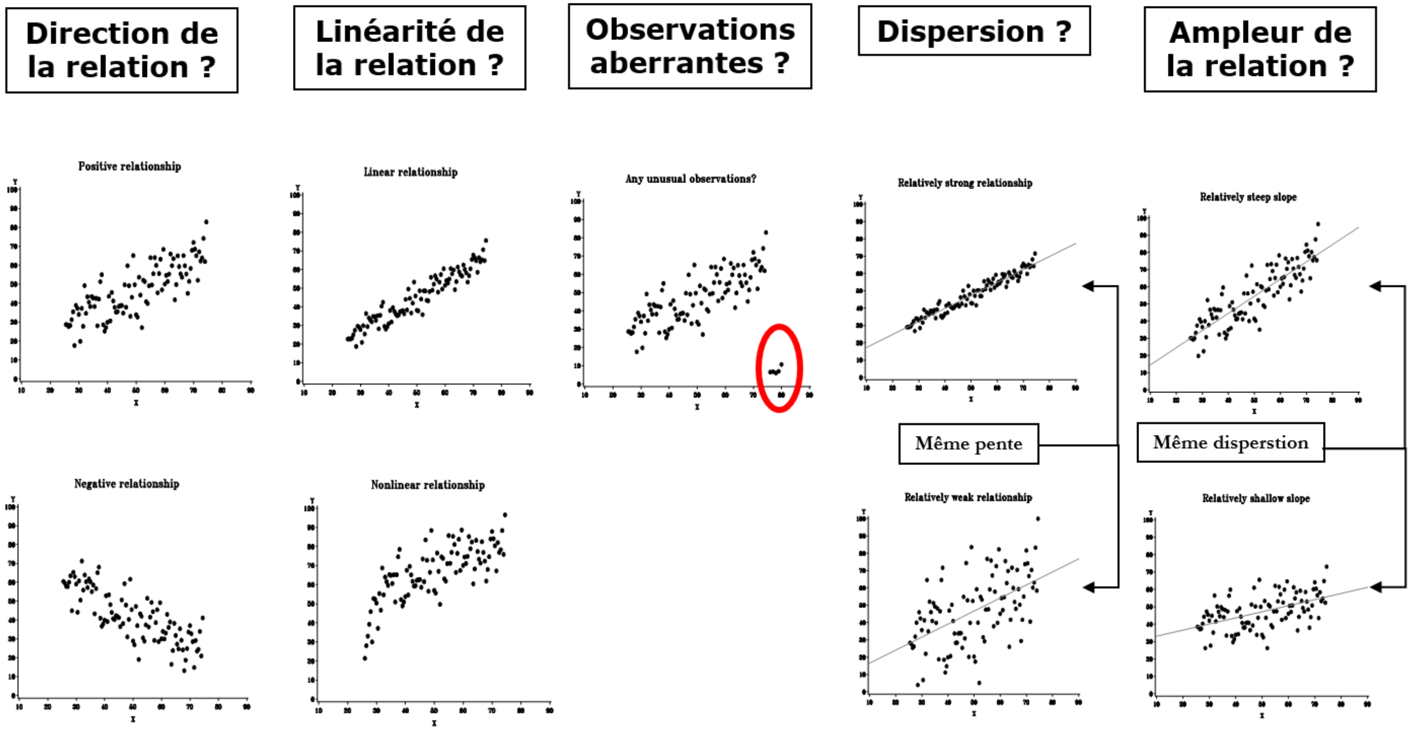
Deux solutions dans ce type de configurations :

- Trouver une **autre fonction (non linéaire)** qui définisse mieux la relation entre x et y (par ex. polynomiale). Cela dépasse le cadre de DECA2
- On peut sinon tenter d'identifier **deux segments linéaires** au sein de la distribution
- La droite rouge caractérise le début de la distribution des x (PIB/hab<5000$)
- La droite verte la fin de la distribution des x (PIB/hab>5000$)



En résumé

Quoi interpréter sur notre nuage de points?



Le coefficient de corrélation

Une autre manière pour savoir si deux variables sont fortement corrélées ou non et connaître le sens de cette corrélation, est de calculer le **coefficient de corrélation linéaire**, que l'on appelle **r de Pearson**

Suppose au préalable d'avoir observé le nuage de points pour savoir **si l'hypothèse de linéarité est acceptable**

Formule du r de Pearson **:**

$$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$ **ou**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

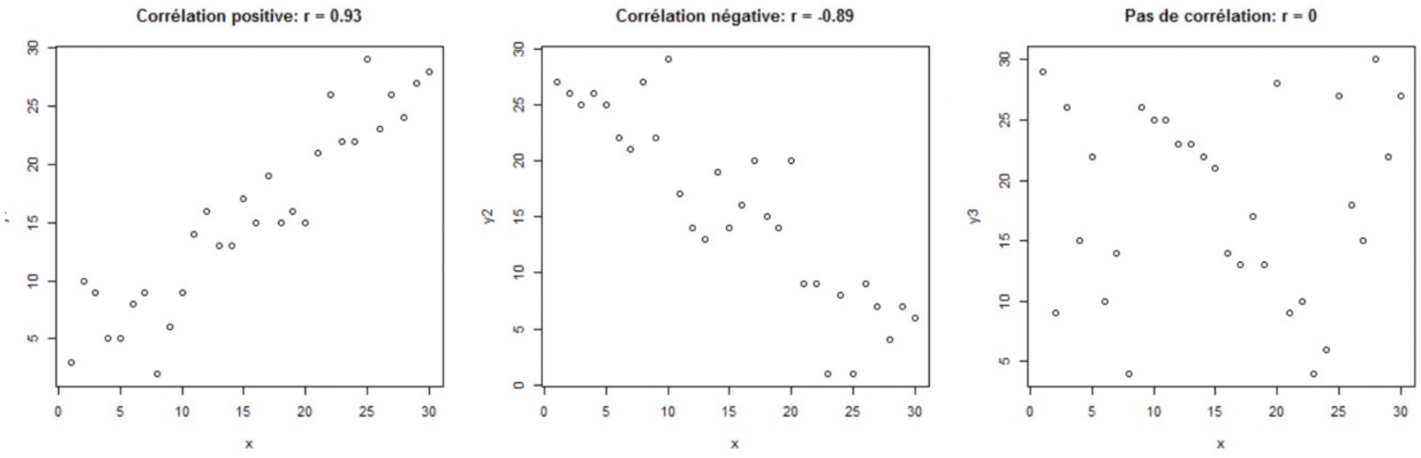
Symbole	**Interprétation**
$\sum$	Somme
σ	Ecart-type
x_i ; y_i	Valeurs réelles
$\bar{x}_i$; $\bar{y}_i$	Valeurs moyennes
Cov	Covariance

Le coefficient de corrélation (2)

Par construction, le coefficient de corrélation fluctue **entre -1 et 1**

Deux éléments sont à interpréter dans la valeur du coefficient :

- Son **signe** : indique le **sens de la relation** entre les deux variables (positive ou négative)
- Sa **valeur** : indique l'**intensité de la relation** ; plus r est éloigné de 0 (0 = absence de corrélation), plus l'intensité de la relation est forte



Le coefficient de corrélation (3)

On peut observer ici comment varie le **r de Pearson** lorsque la distribution se modifie : [Understanding Correlations | R Psychologist]

Quel est selon vous le **r de Pearson** entre PIB/hab et mortalité infantile :

- Pour les pays dont le PIB/hab<5000$ (droite rouge, retour diagramme)?
  - Réponse : -0,68
- Pour les pays dont le PIB/hab est compris entre 5000 et 20000$ (droite verte)
  - Réponse : -0,36
- Pour ces mêmes pays sans la Guinée Equatoriale (**outlier**) ?
  - Réponse : -0,45

Lecture du r de Pearson :

Corrélation	Force	Direction
-1,0 à -0,9	Très fort	Négatif
-0,9 à -0,7	Fort	Négatif
-0,7 à -0,4	Modéré	Négatif
-0,4 à -0,2	Faible	Négatif
-0,2 à 0	Négligeable	Négatif
0 à 0,2	Négligeable	Positif
0,2 à 0,4	Faible	Positif
0,4 à 0,7	Modéré	Positif
0,7 à 0,9	Fort	Positif
0,9 à 1,0	Très fort	Positif

Le coefficient de corrélation (4)

Les limites du coefficient de corrélation linéaire de Pearson :

- Est conçu pour estimer une corrélation **linéaire,** ce qui justifie de regarder au préalable la forme de la distribution par un nuage de points
- Il est dans tous les cas totalement inadapté si la relation n'est pas **monotone** (parfois croissante, parfois décroissante)
- Le coefficient de corrélation est sensible aux **outliers**
- Le coefficient n'indique en rien le **sens de la causalité** entre les deux variables (par construction, intervertir x et y dans la formule amène au même résultat)

Exercice d’application sur Jamovi