



---

# THÉMATIQUE 2 - ANALYSE BIVARIÉE : CROISEMENT QUANTITATIF-QUANTITATIF

---

Data Sciences Revision  
MENTION MATHÉMATIQUES ET INFORMATIQUE  
PARCOURS HPDA

12/09/2024

*Rédigé par :*  
PAULY ALEXANDRE  
alexandre.pauly@cy-tech.fr

L'objectif d'une analyse bivariee est de faire lien entre deux variables (ici quantitatives). Il est donc possible de faire le lien entre le prix de voiture et le nombre de vente de cette voiture par exemple.

## 1 Analyser les données

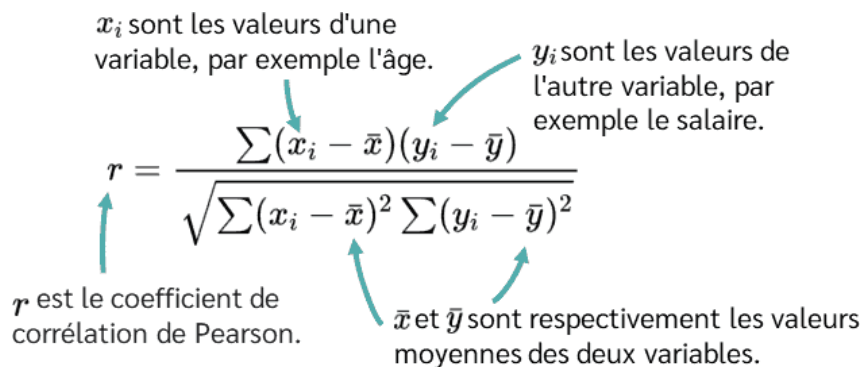
### 1.1 Coefficient de corrélation

Afin de quantifier l'intensité du **lien statistique** qui peut exister entre deux variables, nous allons utiliser le coefficient de corrélation.

Cet indice est compris entre -1 et 1. Plus il est proche de  $|1|$ , plus le lien est sensé être fort :

- +1 : Les deux variables croissent ou décroissent conjointement
- -1 : Lorsqu'une variable croît, l'autre décroît
- 0 : Pas de relation entre les deux variables

**Attention** : Une liaison, même très forte, entre deux variables, n'indique pas la causalité!



The diagram shows the formula for the Pearson correlation coefficient  $r$  with several annotations in French:

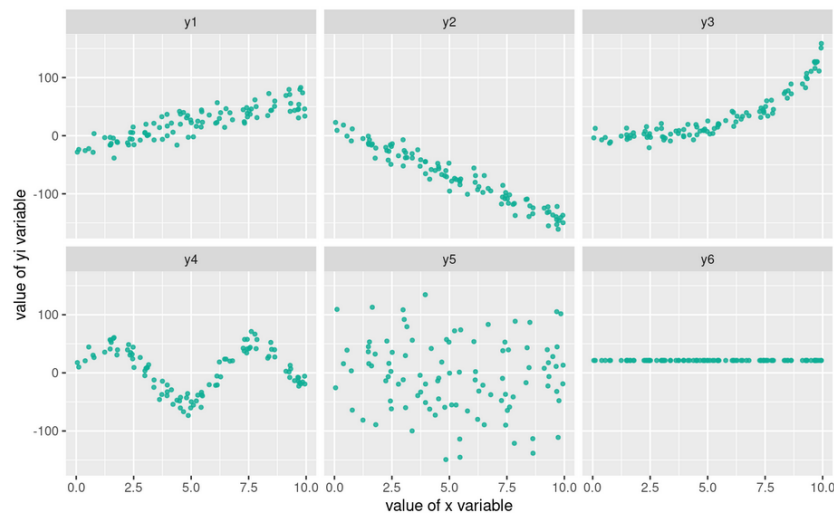
- An arrow points from the text " $x_i$  sont les valeurs d'une variable, par exemple l'âge." to the  $x_i$  term in the numerator.
- An arrow points from the text " $y_i$  sont les valeurs de l'autre variable, par exemple le salaire." to the  $y_i$  term in the numerator.
- An arrow points from the text " $\bar{x}$  et  $\bar{y}$  sont respectivement les valeurs moyennes des deux variables." to the  $\bar{x}$  and  $\bar{y}$  terms in the denominator.
- An arrow points from the text " $r$  est le coefficient de corrélation de Pearson." to the  $r$  on the left side of the equation.

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

Figure 1: Coefficient de corrélation

### 1.2 Diverses formes de dépendances

Le lien entre les variables expriment le fait que les valeurs de ces variables n'évoluent pas indépendamment, mais au contraire présentent une certaine forme/logique.



**Figure 2:** Coefficient de corrélation

La Figure ci-dessus (**figure 2** montre certaines relations que l'on peut obtenir :

- linéaire positive (y1)
- linéaire négative (y2)
- non-linéaire, peut-être exponentielle (y3)
- périodique, sinusoïdale (y4)
- absence de dépendance, les deux variables sont indépendantes (y5)
- absence de dépendance, la variable de l'axe des y est constante (y6)

**Attention :** Certains jeux de données peuvent contenir des outliers, des données bruitées, etc.

### 1.3 Les étapes à vérifier

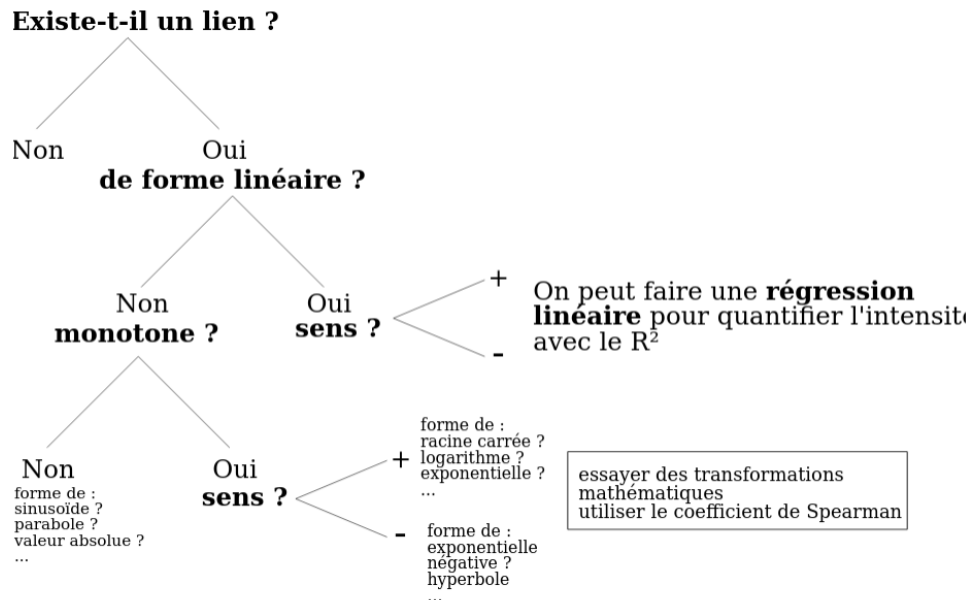


Figure 3: Etapes à suivre pour analyser les données

## 2 Régression linéaire

Après avoir regardé l'aspect des données avec un graphique, il est logique d'entamer une régression linéaire si le nuage de points forme allongée. Dans le cas contraire, le résultat peut être inutile.

### 2.1 Utiliser un modèle linéaire

En python, de nombreux modules permettent de réaliser une régression linéaire (Scikit-learn, stats, lineregress, etc.). Parfois, il faudra effectuer certaines transformations pour adapter un jeu de données à un cas linéaire (utilisation de log, exp, cos, etc.).

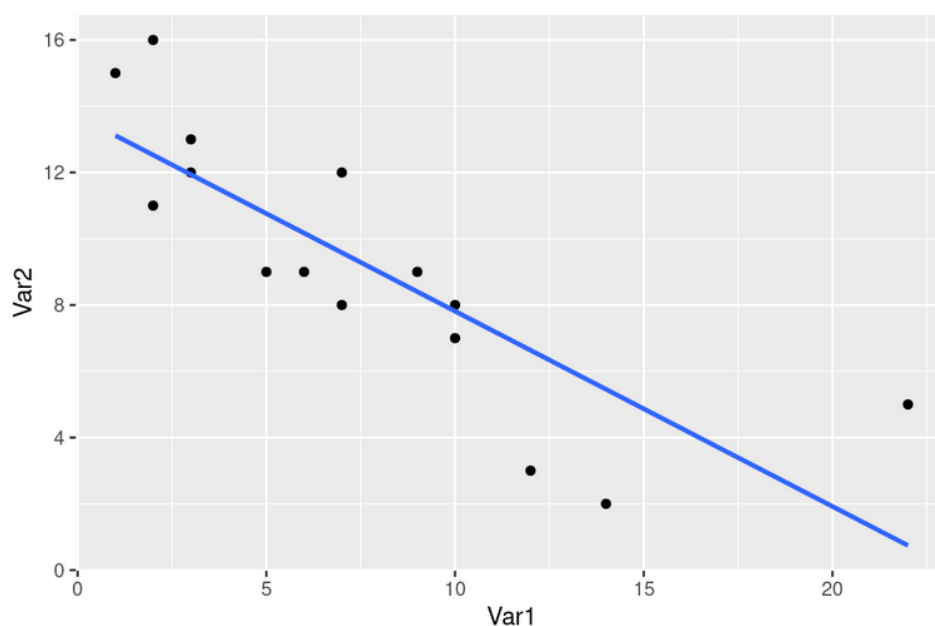
Famille	Fonctions	Transformation
exponentielle	$y = a.e^{bx}$	$y' = \log(y)$
puissance	$y = ax^b$	$y' = \log(y) \quad x' = \log(x)$
inverse	$y = a + \frac{b}{x}$	$x' = \frac{1}{x}$
logistique	$y = \frac{1}{1 + e^{-(a.x+b)}}$	$y' = \log\left(\frac{y}{1-y}\right)$

**Figure 4:** Transformations pour une régression linéaire

## 2.2 Analyser une régression

L'objectif de la régression linéaire est trouver le meilleur modèle linéaire entre deux variables. Ce modèle est l'équation d'une droite, qu'on appelle droite de régression et qui permet de visualiser :

- l'intensité de la dépendance, suivant que les points sont proches de la droite ou non
- la forme de la dépendance, suivant que le nuage soit bien de forme linéaire
- le sens de la dépendance : nulle, positive ou négative



**Figure 5:** Régression linéaire

## 2.3 Evaluation

Le coefficient de détermination ( $R^2$ ) permet de s'assurer qu'une régression linéaire est de bonne qualité. Lorsque le  $R^2$  est élevé (proche de 1), le modèle prédit "bien" les observations. De même, si les p-values associées sont faibles, il y a "peu" de chances de se tromper.

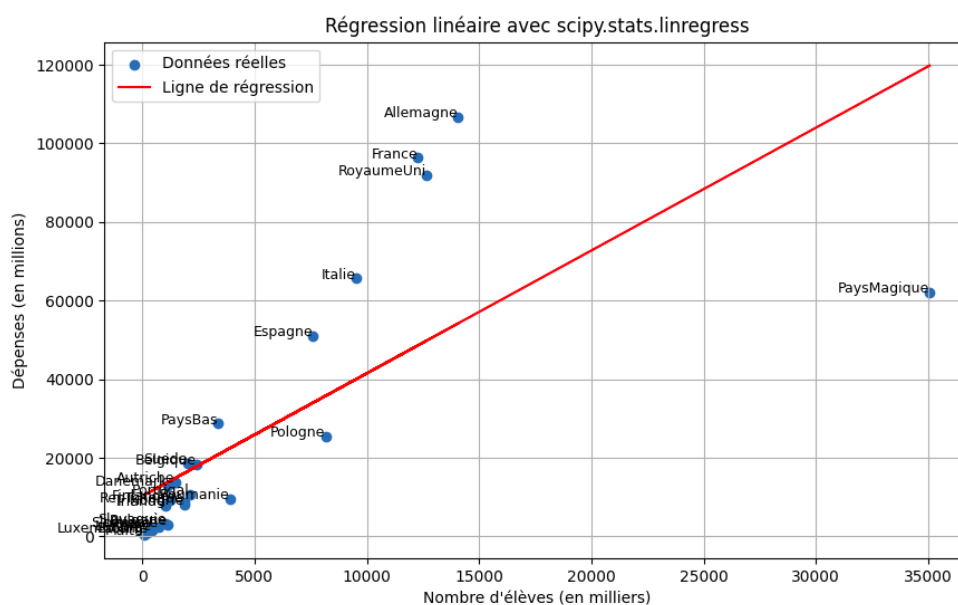
## 2.4 Prédiction

Une fois le modèle établi, les coefficients de la régression permettent de prédire de nouvelles valeurs.

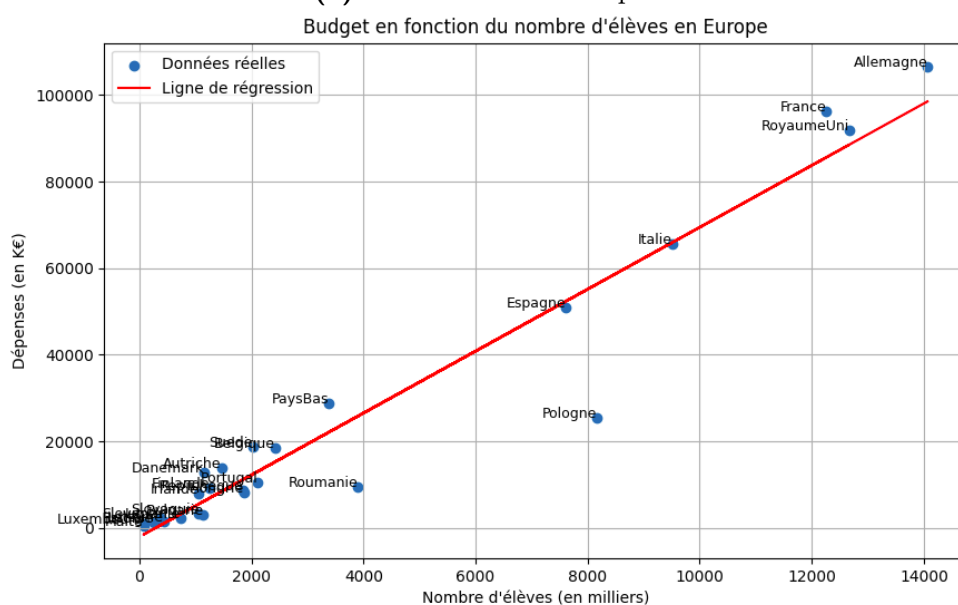
**Attention :** En cas de transformation, ne pas oublier à faire une transformation inverse pour comparer le résultat au jeu de données.

# 3 Gestion des individus atypiques

Les jeux de données comprenant tout un tas d'observations, il est parfois difficile d'approcher avec fiabilité un dataset sans l'élaguer. Pour cela, il est utile de vérifier les hypothèses sur les résidus. Ceci s'analyse facilement avec la figure suivante qui met en lumière l'écart d'un individu vis-à-vis du reste du groupe.



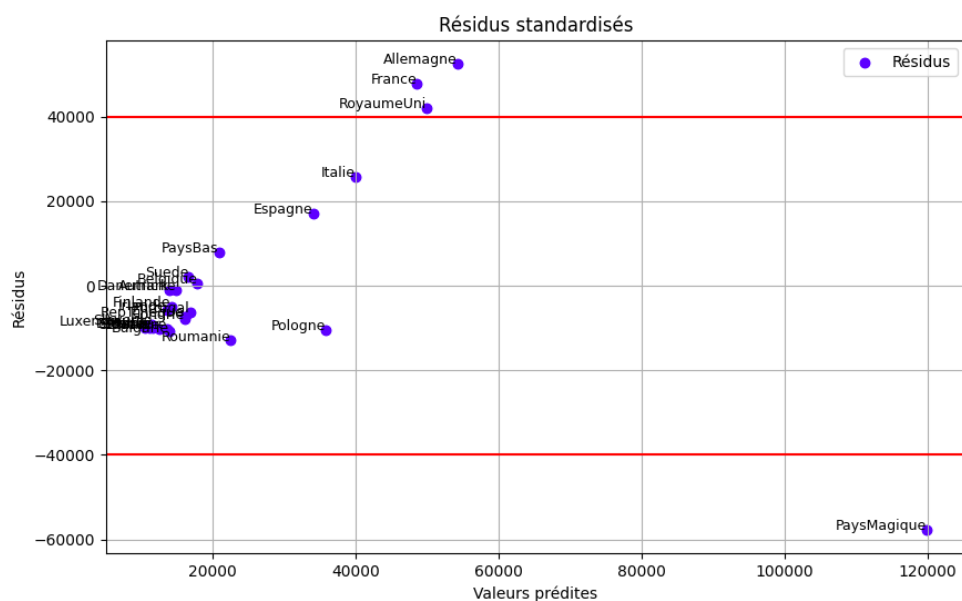
(a) Jeu de données complet



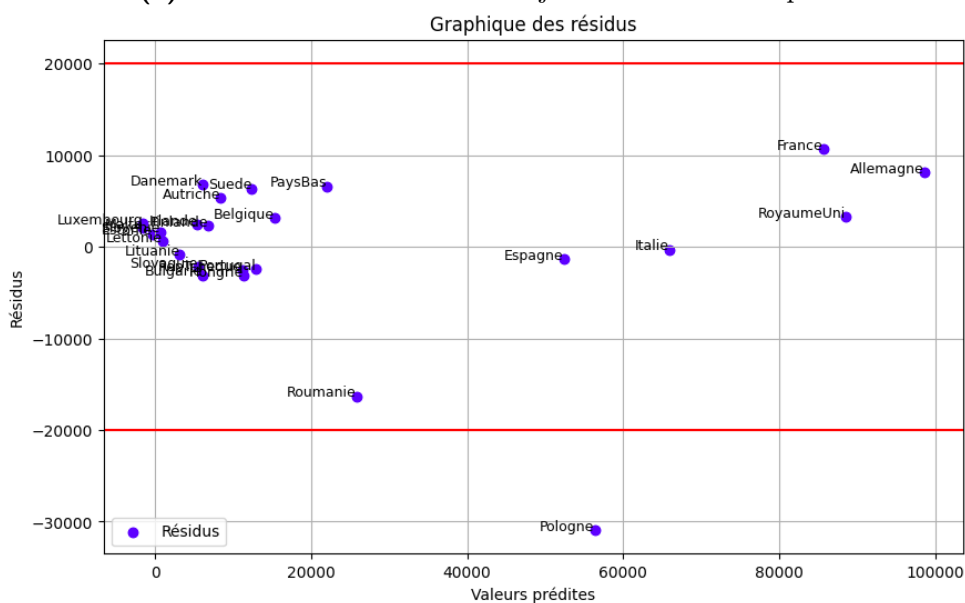
(b) Jeu de données sans l'individu atypique 'PaysMagique'

Figure 6

Comme le montre la figure ci-dessus, le modèle approche bien mieux les données sans l'individu 'PaysMagique'.



(a) Résidus standardisés sur le jeu de données complet



(b) Résidus standardisés sur le jeu de données sans l'individu atypique 'PaysMagique'

**Figure 7:** Visualisation des résidus normalisés

Comme la montre le (b) de la **figure 6**, maintenant que 'PaysMagique' a été retiré, l'individu 'Pologne' est lui aussi considéré comme atypique sur cette nouvelle modélisation.



## References

- [1] Chapitre 5 Analyse Bivariée | Analyse Statistique M2 IGAST, et DESIGEO. (s. d.). LASTIG. [https://www.umr-lastig.fr/paul-chapron/resources/cours\\_site/bivariee.htmlcontenu-du-chapitre](https://www.umr-lastig.fr/paul-chapron/resources/cours_site/bivariee.htmlcontenu-du-chapitre), [https://www.umr-lastig.fr/paul-chapron/resources/cours\\_site/bivariee.htmlcontenu-du-chapitre](https://www.umr-lastig.fr/paul-chapron/resources/cours_site/bivariee.htmlcontenu-du-chapitre).