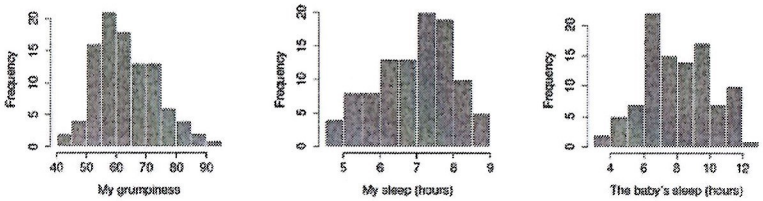
Statistiques bivariées - Analyse de graphique

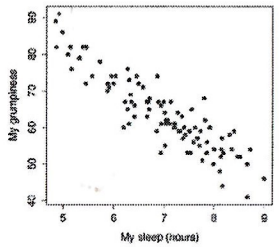
Exemple illustratif

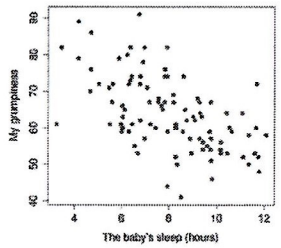
Supposons 3 variables : Grinchiosité de Dan (*dan.grump*), Les heures de sommeil de Dan (*dan.sleep*), Les heures de sommeil du fils de Dan (*baby.sleep*), distribuées de la façon suivante (analyse univariée) :



Nous pouvons dessiner des diagrammes de dispersion (ou nuage de points) pour nous donner une idée générale du degré de corrélation entre deux de ces variables. Par exemple, comparons la relation entre dan.sleep et dan.grump (1a) et celle entre baby.sleep et dan.grump (1b).

Graph 1a



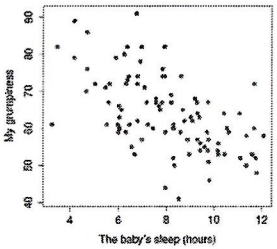


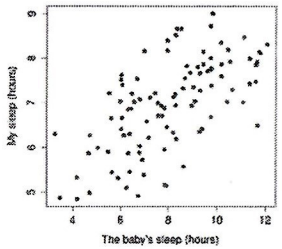
Graph 1b

En regardant ces deux figures côte à côte, il est clair que la relation est qualitativement la même dans les deux cas : plus de sommeil égal moins de grinchiosité ! Cependant, il est aussi assez évident que la relation entre dan.sleep et dan.grump est plus forte que la relation entre baby.sleep et dan.grump. La figure de gauche est « plus nette » que celle de droite. Il semble que si vous vouliez prédire l'humeur de Dan, cela vous aiderait un peu de savoir combien d'heures son fils a dormi, mais ce serait plus utile de savoir combien d'heures il a dormi lui.

En revanche, si l'on compare le nuage de points de baby.sleep/dan.grump (2a) au nuage de points de baby.sleep/dan.sleep (2b), la force globale de la relation est la même, mais la direction est différente. C'est-à-dire, si son fils dort plus, Dan dort plus (relation positive, 2b), et si son fils dort plus, Dan est moins grincheux (relation négative, 2a).

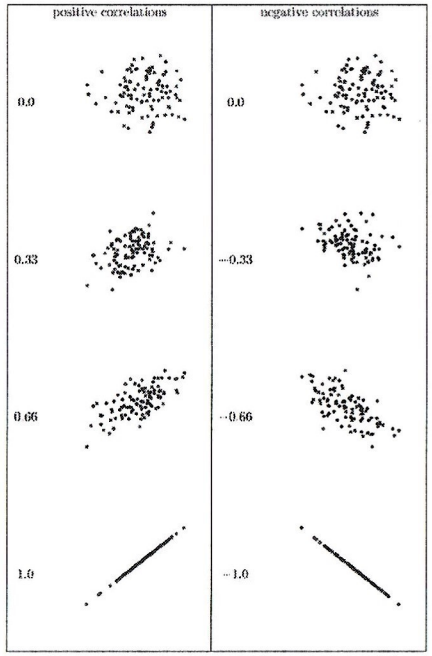
Graph 2a





Graph 2b

Lien avec le coefficient de corrélation linéaire



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

Guides approximatifs pour interpréter les corrélations

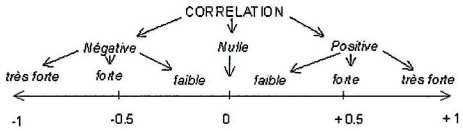
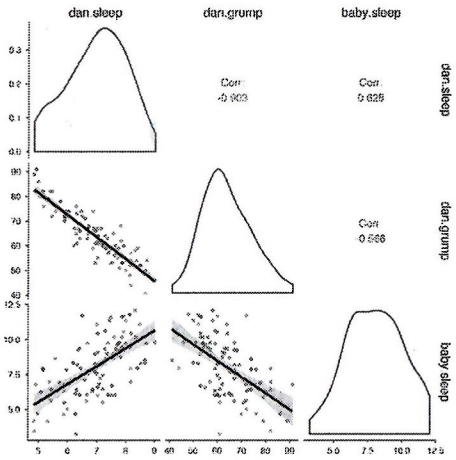


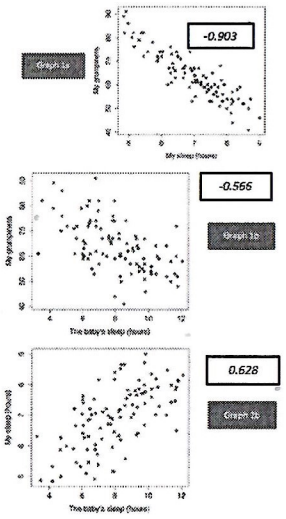
Illustration de l'effet de la variation de l'intensité et de la direction d'une corrélation. Les chiffres mentionnés correspondent aux coefficients de corrélation linéaire.

Dans nos exemples précédents, les coefficients linéaires des graph 1a et 1b (et donc 2a) sont négatifs et celui du graph 2b est positif. En terme d'intensité, la valeur absolue du coefficient du graph 1a est plus élevée que celle des autres graphs.

En effet, voici la matrice correspondant aux croisements de ces 3 variables :

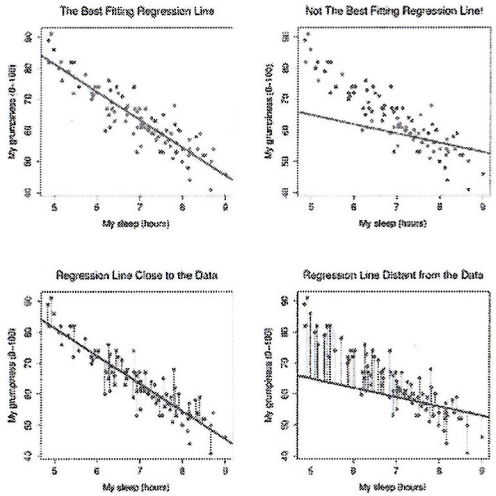


D'après cette matrice, nous voyons que les coefficients de corrélation linéaire des graphs précédents sont les suivants :



Lien avec le modèle de régression linéaire

Les modèles de régression linéaire, réduits à l'essentiel, sont essentiellement une version légèrement plus sophistiquée du coefficient de corrélation linéaire. Nous avons dessiné quelques nuages de points pour nous aider à examiner la relation entre la quantité de sommeil obtenue par Dan et sa mauvaise humeur le lendemain, et cela correspondait à une corrélation linéaire de r = -0,903. Mais quand nous avons regardé le graphique, ce que nous nous représentions mentalement était une ligne droite au milieu des données. En statistique, cette ligne que nous traçons s'appelle une ligne de régression. Notez qu'instinctivement, la ligne de régression imaginée passe au milieu des données.



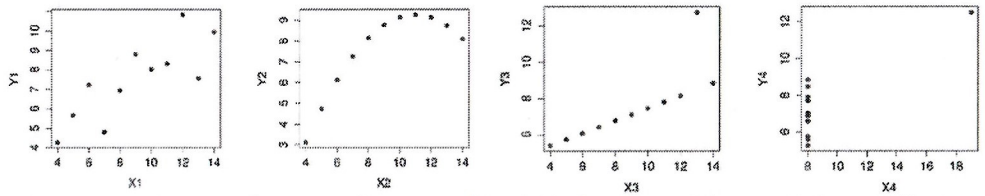
La figure de gauche montre le diagramme de dispersion du sommeil Graph 1a avec la ligne de régression la mieux adaptée tracée au-dessus du sommet. Comme on pouvait s'y attendre, la ligne passe au milieu des données. En revanche, la figure de droite montre les mêmes données, mais avec un très mauvais choix de ligne de régression tracée par-dessus.

Rappelons que la méthode des moindres carrés ordinaires permet de définir l'ajustement linéaire tel que la somme de toutes les distances entre chaque point du graphique et la droite (au carré) soit la plus petite possible. Si nous ajoutons les écarts, les distances, de chaque point à la droite, aux graphiques précédents, nous obtenons les graphiques ci-contre :

Nous voyons que ces écarts sont beaucoup plus petits et moins importants pour la bonne ligne de régression. La somme de tous ces écarts (au carré) sera donc plus petite à gauche qu'à droite.

Enfin, notons l'importance de l'analyse graphique en premier lieu : **il faut toujours commencer par tracer ses données** !!!

Le quatuor d'Anscombe est l'exemple le plus connu :

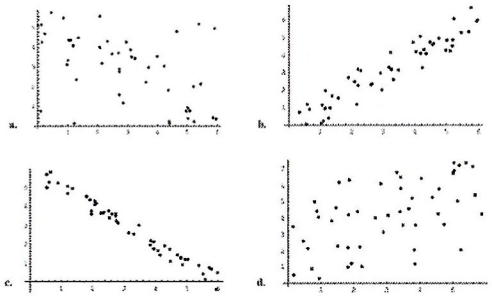


Ces quatre ensembles de données ont un coefficient de corrélation linéaire de r = 0,816, mais ils sont qualitativement différents les uns des autres. Pourtant, pour les quatre jeux de données, la valeur moyenne pour X à 9 et la moyenne pour Y à 7,5. Les écarts-types pour toutes les variables X sont presque identiques, tout comme ceux des variables Y.

Exercice

Rendez à chacun des nuages de points ci-contre son coefficient de corrélation linéaire :

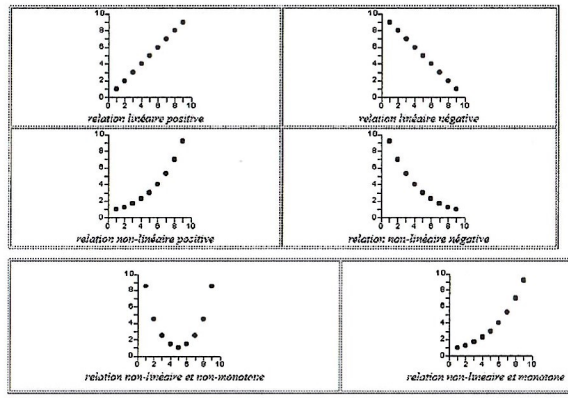
−0.98 −0.50 0.53 0.94



Ajustement non-linéaire

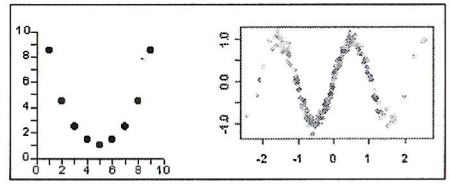
Dans certains cas, l'ajustement à une fonction linéaire n'est pas adéquat : un ajustement des données à une fonction non linéaire doit être envisagé.

Une relation non-linéaire est monotone si elle est strictement croissante ou strictement décroissante, c'est-à-dire si elle ne comporte pas de minima ou de maxima. Toutes les relations linéaires sont monotones.

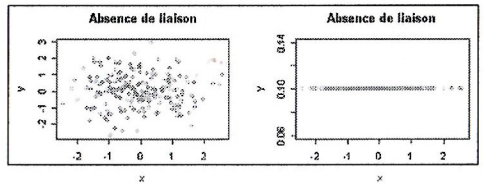


Pour les graphiques représentant des relations non-linéaires, le coefficient de corrélation linéaire ne sera pas un bon indicateur. En effet, il sera modéré voire fort mais l'approximation linéaire de cette relation sera erronée. Toutefois, ces deux variables présenteront un fort rapport de corrélation η, qui ne présage rien de la forme de la relation.

Enfin, d'autres relations non linéaires peuvent exister, sur lesquelles nous pouvons conclure à tort si nous ne regardons que le coefficient de corrélation linéaire : les relations non monotones. Les graphiques de relations non monotones présentent de forts rapports de corrélation et de faibles coefficients de corrélation linéaire. Les graphiques ci-contre en sont des exemples. Nous voyons bien ici que ces graphiques peuvent difficilement être approchés par une droite. Pourtant, il existe une relation entre les variables.



Pour rappel, une absence de relation se traduit par un graphique présentant des points diffus ou sans variation :



Exemple de coefficient de corrélation linéaire en fonction de différents nuages de points