Cours de biostatistique ⋈ Illustrations dans 

# De la stature chez l'Homme ... à la taille des cerveaux chez les mammifères. Réversion, Régression, Corrélation

A.B. Dufour, J.R. Lobry & D. Chessel

31 mars 2008

## Table des matières

# 1 Un peu d'histoire ...

En 1895, Karl Pearson (1857-1936) donnait la paternité de la corrélation à Auguste Bravais (1811-1863) puis il revint sur cet avis en 1920. Les historiens de la statistique [4] s'entendent aujourd'hui pour dire que le rôle essentiel a été joué par Francis Galton (1822-1911).

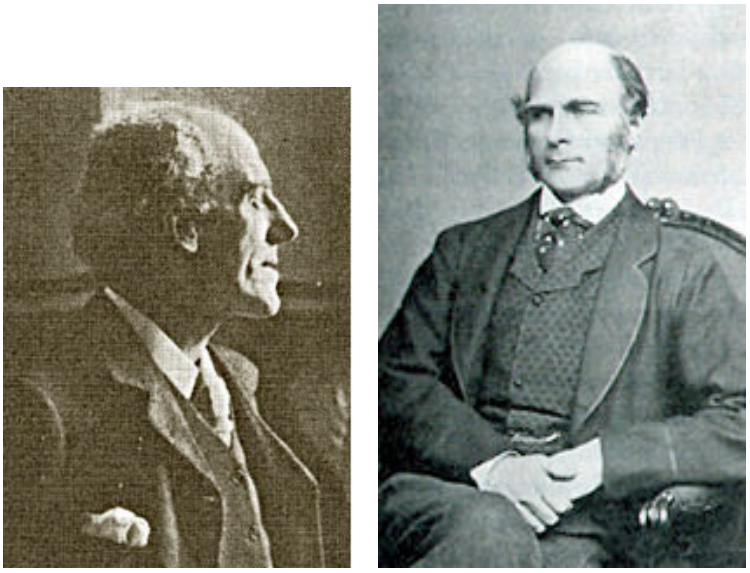


FIG. 1 – K. Pearson et F. Galton

Ce dernier a souligné la nécessité d'une mesure de la corrélation dans l'analyse des séries bivariées. Et c'est par le concept de régression qu'il débute. Le 9 février 1877, Galton fait un exposé à l'Institution Royale de Grande-Bretagne, intitulé *Typical laws of heredity in man.*

> *Reversion is a tendency of the ideal mean filial type to depart from the parental type, reverting to what may be roughly and perhaps fairly described as the average ancestral type. If family variability has been the only process in simple descent that affected the characteristics of a sample, the dispersion of the race from its mean ideal type would indefinitely increase with the number of generations, but reversion checks this increase, and brings it to a standstill.*

Galton exprime le désir de construire un coefficient de réversion qui indique la réduction de la variabilité de la famille. Cette réversion se transforme peu à peu en régression puis en corrélation [6].

> *It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were wholly due to common causes, the co-relation would be perfect, as in approximately the case with the symmetrically disposed parts of the body. If they were in no respect due to the common causes, the co-relation would be nil. Between these two extremes are an endless*

*number of intermediate cases and it will be shown how the closeness of co-relation in any particular case admits of being expressed by a singular number.*

C'est en 1896 que Karl Pearson [8] reprend le concept et lui donne la forme que nous connaissons aujourd'hui.

L'objectif de cette séance est de définir, comprendre la nature du coefficient de corrélation linéaire et de lier la représentation graphique associée au croisement de deux variables quantitatives appelée *nuage de points* et ce fameux coefficient.

Analysons les données [6] de F. Galton (1886) sur la relation entre la taille (en pouces : 1 pouce vaut 2.54 cm) de 928 enfants et la taille de leurs parents (en pouces). Comme un enfant a deux parents, Galton a traité le problème en introduisant la notion du "mid-parent" en prenant la moyenne de la taille du père avec la taille de la mère multipliée par 1.08. Ce coefficient a été construit à partir de la moyenne des tailles des pères et la moyenne des tailles des mères. Les données ont été trouvées sur le site internet : `http://www.mugu.com/galton/index.html`. Importez les dans R avec la commande suivante :

```
taimdc <- read.table("http://pbil.univ-lyon1.fr/R/donnees/taimdc.txt",
    header = TRUE)
summary(taimdc)
```

```
       x               y        
 Min.   :64.00   Min.   :61.70  
 1st Qu.:67.50   1st Qu.:66.20  
 Median :68.50   Median :68.20  
 Mean   :68.31   Mean   :68.09  
 3rd Qu.:69.50   3rd Qu.:70.20  
 Max.   :73.00   Max.   :73.70  
```

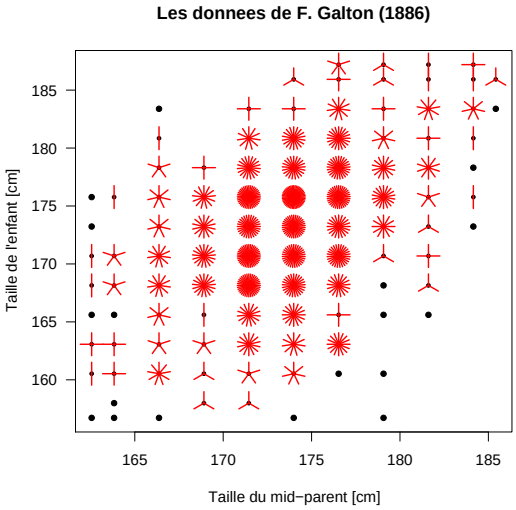
Les données sont en pouces, ce n'est pas très lisible, convertissez les en centimètres :

```
taimdc <- 2.54 * taimdc
summary(taimdc)
```

```
       x               y        
 Min.   :162.6   Min.   :156.7  
 1st Qu.:171.4   1st Qu.:168.1  
 Median :174.0   Median :173.2  
 Mean   :173.5   Mean   :172.9  
 3rd Qu.:176.5   3rd Qu.:178.3  
 Max.   :185.4   Max.   :187.2  
```

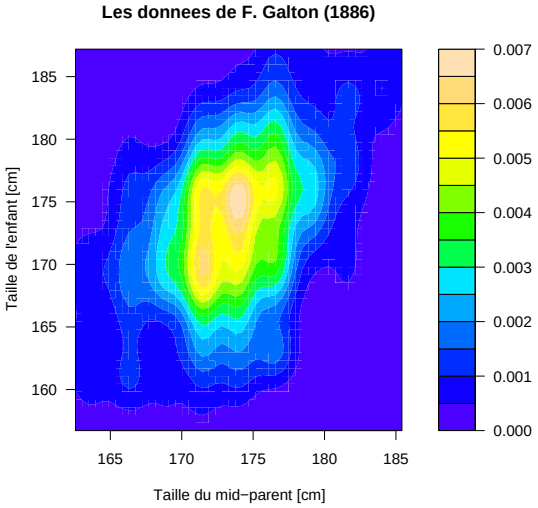
Représentez les graphiquement :

```
sunflowerplot(taimdc, xlab = "Taille du mid-parent [cm]", ylab = "Taille de l'enfant [cm]",
    las = 1, main = "Les donnees de F. Galton (1886)")
```




Nous reconnaissons dans cette figure la représentation graphique appelée "sunflowerplot". Nous avons d'une part les points correspondant au croisement des tailles "parents, enfants" et les branches décomptent le nombre de répétitions. On peut utiliser également une représentation avec une carte de densité :

```
library(MASS)
densite <- kde2d(taimdc$x, taimdc$y, n = 50)
filled.contour(densite, color = topo.colors, xlab = "Taille du mid-parent [cm]",
    ylab = "Taille de l'enfant [cm]", las = 1, main = "Les donnees de F. Galton (1886)")
```




Le coefficient de corrélation calculé sur ces données est égal à 0.46.

```
cor(taimdc$x, taimdc$y)
[1] 0.4587624
```

Mais que signifie-t-il ?

# 2 La corrélation

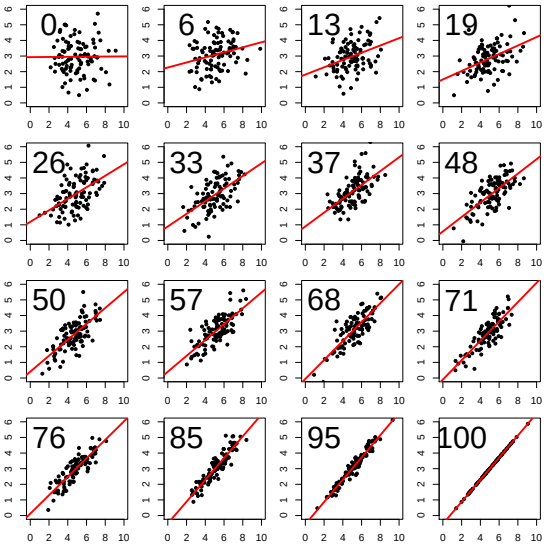
## 2.1 Qu'est-ce que c'est ?

Le *coefficient de corrélation linéaire* est une mesure de la liaison linéaire, c'est-à-dire de la capacité de prédire une variable $x$ par une autre $y$ à l'aide d'un modèle linéaire. Pour ce faire, profitons du logiciel R pour réaliser quelques simulations.
Allez dans votre dossier de travail, puis ouvrez un document texte seul (à l'aide du clic droit de la souris). Nous allons construire les fonctions suivantes : `simu()` crée des nuages de points, `nuage()` construit ces nuages en superposant le modèle linéaire.

```
simu <- function(rho) {
    n = 100
    for (i in 1:100) {
        cov <- matrix(c(2, sqrt(2) * rho, sqrt(2) * rho, 1), 2,
            2)
        w <- mvrnorm(n, c(5, 3), cov)
        w <- data.frame(w)
        names(w) <- c("x", "y")
        robs <- cor(w$x, w$y)
        if (((rho - 0.02) < robs) & (robs < (rho + 0.02)))
            return(w)
    }
    return(w)
}
nuage <- function(X) {
    plot(X$x, X$y, xlim = c(0, 10), ylim = c(0, 6), pch = 20)
    lm0 <- lm(y ~ x, data = X)
    abline(lm0, col = "red", lwd = 2)
    r2 <- summary(lm0)$r.squared
    text(2, 5, as.integer(100 * r2), cex = 3)
}
```

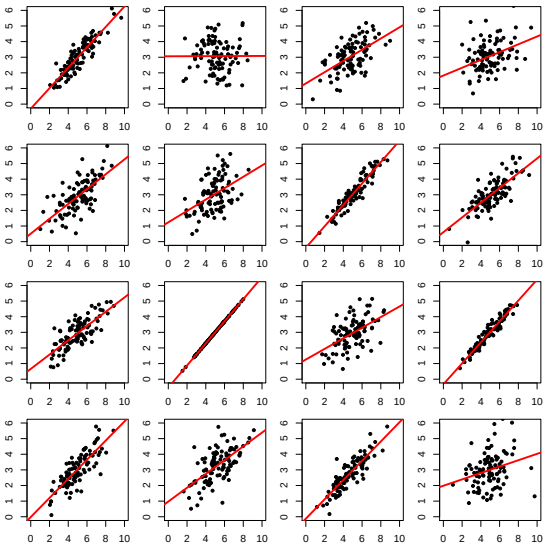
Maintenant, allez sous R puis dans le menu file, Source R code. Dans la fenêtre "Select file to source", sélectionner tous les types de fichiers puis ouvrir le fichier ".txt" comprenant la fonction que vous venez de créer. Il se peut que des messages d'erreurs soient signalés. Cela signifie que vous avez mal rentré le "texte". Ré-itérer l'opération jusqu'à ce qu'il n'y ait plus d'erreur. Puis taper les commandes suivantes :

```
par(mfrow = c(4, 4))
par(mar = c(2, 2, 1, 1))
w <- lapply(sqrt(seq(0, 1, le = 16)), simu)
lapply(w, nuage)
```




On a un nombre compris entre 0 et 1 (noté ici en pourcentage sur les graphes) qui mesure la qualité de la prédiction de $y$ par $x$ à l'aide d'une droite (en rouge). C'est le carré de la corrélation.

**Exercice.** Les graphiques suivants :



correspondent aux 16 carrés de corrélations suivants :

```
 [1] 0.000 0.072 0.146 0.206 0.269 0.317 0.396 0.471 0.562 0.595 0.668 0.763 0.823
[14] 0.892 0.932 1.000
```

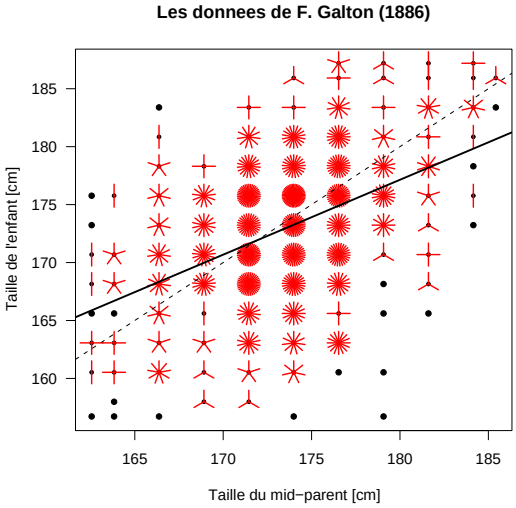
Qui est qui ?

## 2.2 Comment calculer la corrélation ?

Il faut mettre dans le nuage une **droite** qui s'ajuste au mieux puis mesurer la qualité de cet ajustement optimal. Le critère est celui des moindre carrés.

Reprendre les données de Galton et appliquez les fonctions ci-dessous.

```
sunflowerplot(taimdc, xlab = "Taille du mid-parent [cm]", ylab = "Taille de l'enfant [cm]",
    las = 1, main = "Les donnees de F. Galton (1886)")
abline(lm(taimdc$y ~ taimdc$x), lwd = 2)
abline(c(0, 1), lty = 2)
```

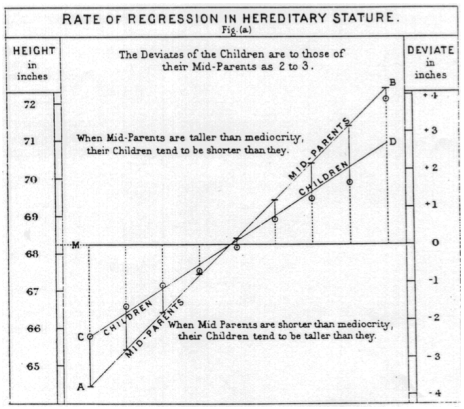



Les coefficients de la droite sont donnés par :

```
coefficients(lm(taimdc$y ~ taimdc$x))
```

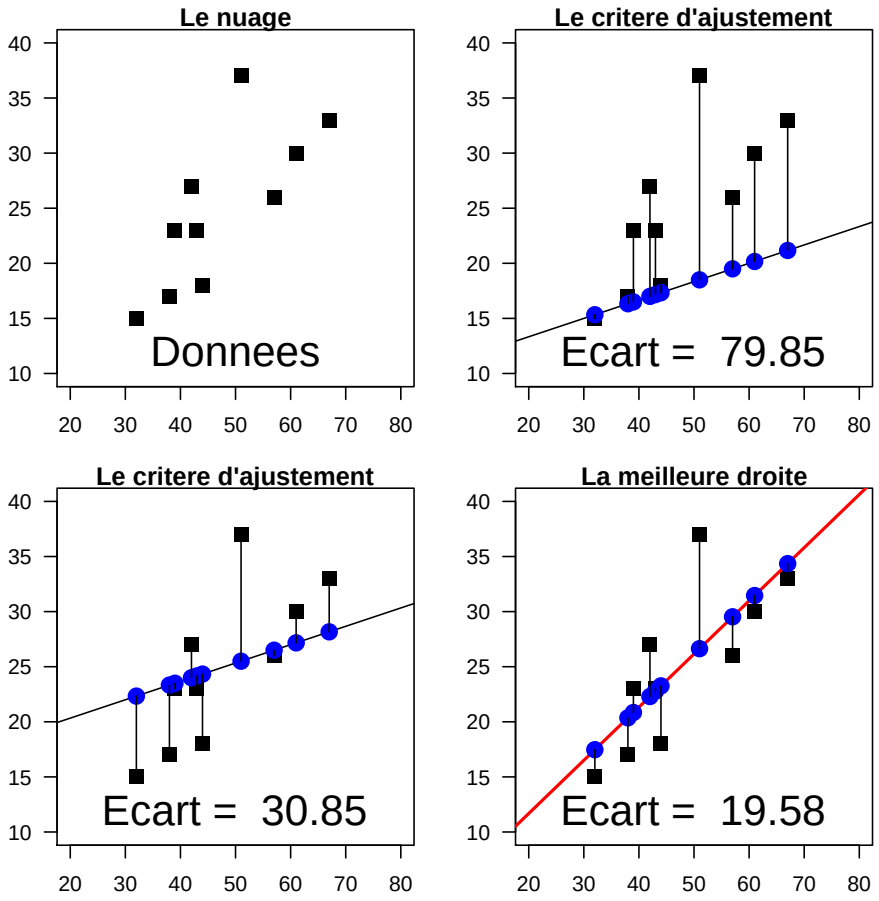
```
(Intercept)     taimdc$x
 60.8114867    0.6462906
```

On constate que la pente de la droite est inférieure à un (sur le dessin précédent la droite en pointillés est d'équation $y = x$). C'est l'origine historique du terme de droite de *régression* : les enfants de parents de grande taille ont tendance à être plus petits qu'eux, les enfants de parents de petite taille ont tendance à être plus grands qu'eux. Galton parlait de régression vers la médiocrité :




Le terme est resté. Illustrons le principe sur un petit jeu de données fictives :

```
x <- c(61, 67, 32, 43, 57, 44, 39, 38, 42, 51)
y <- c(30, 33, 15, 23, 26, 18, 23, 17, 27, 37)
par(mfrow = c(2, 2))
par(mar = c(3, 3, 1, 1))
xlim <- c(20, 80)
ylim <- c(10, 40)
plot(x, y, xlim = xlim, ylim = ylim, pch = 15, cex = 1.5, las = 1,
    main = "Le nuage")
text(50, 12, "Donnees", cex = 2)
plot(x, y, xlim = xlim, ylim = ylim, pch = 15, cex = 1.5, las = 1,
    main = "Le critere d'ajustement")
ytheo <- x/6 + 10
abline(c(10, 1/6))
points(x, ytheo, pch = 20, col = "blue", cex = 2.5)
segments(x, ytheo, x, y)
mess <- paste("Ecart = ", round(mean((y - ytheo)^2), dig = 2))
text(50, 12, mess, cex = 2)
plot(x, y, xlim = xlim, ylim = ylim, pch = 15, cex = 1.5, las = 1,
    main = "Le critere d'ajustement")
ytheo <- x/6 + 17
abline(c(17, 1/6))
points(x, ytheo, pch = 20, col = "blue", cex = 2.5)
segments(x, ytheo, x, y)
mess <- paste("Ecart = ", round(mean((y - ytheo)^2), dig = 2))
text(50, 12, mess, cex = 2)
plot(x, y, xlim = xlim, ylim = ylim, pch = 15, cex = 1.5, las = 1,
    main = "La meilleure droite")
lmd <- lm(y ~ x)
ytheo <- predict(lmd)
abline(lmd$coef, col = "red", lwd = 2)
points(x, ytheo, pch = 20, col = "blue", cex = 2.5)
segments(x, ytheo, x, y)
mess <- paste("Ecart = ", round(mean((y - ytheo)^2), dig = 2))
text(50, 12, mess, cex = 2)
```




La droite de régression est celle qui minimise la moyenne des carrés des écarts entre les valeurs observée (`y`) et celles prédites par le modèle (`ytheo`). Le carré du coefficient de corrélation linéaire donne la fraction de la variance de `y` qui est prise en compte par le modèle :

```
 var(ytheo)/var(y)
[1] 0.5734295
 cor(x, y)^2
[1] 0.5734295
```

On prend le signe + si la relation est croissante et le signe - si elle est décroissante. On ajoute la racine de la fraction de la variance expliquée et on a le coefficient de corrélation linéaire :

```
 cor(x, y)
[1] 0.7572513
```

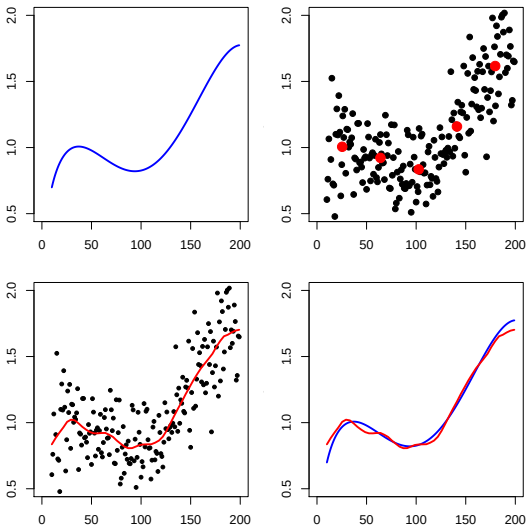
## 2.3 A quoi ça sert ?

Une fonction mathématique inconnue relie une quantité $x$ à une quantité $y$. On veut décrire le phénomène. Il n'est reproductible qu'avec une *erreur d'échantillonnage* forte : on trouve en moyenne le résultat mais un résultat isolé est

entaché d'une forte erreur. Les points de mesure sont régulièrement espacés. On voit le résultat. Localement on fait une régression linéaire. On obtient une estimation en ce point (gros points rouges). On recommence régulièrement ; on obtient une estimation de la fonction. Il est remarquable que la fonction est, en général, assez proche de la fonction de départ. Ces sortes d'outils sont des *lisseurs à régression locale*.

Ecrivez la fonction `fsimul()` dans un fichier texte, 'sourcez' la fonction et réalisez quelques simulations.

```
fsimul <- function() {
    par(mfrow = c(2, 2))
    par(mar = c(3, 3, 1, 1))
    x <- 10:199
    y <- (x - 50) * x * (x - 150) * (230 - x)/1e+08 + log(x)/4
    yobs <- y + rnorm(190, 0, 0.2)
    z <- loess.smooth(x, yobs, span = 0.2)
    plot(x, y, type = "l", lwd = 2, xlim = c(0, 200), ylim = c(0.5,
        2), col = "blue")
    plot(x, yobs, type = "p", pch = 20, xlim = c(0, 200), ylim = c(0.5,
        2), cex = 1.5)
    points(z$x[c(5, 15, 25, 35, 45)], z$y[c(5, 15, 25, 35, 45)],
        cex = 2.5, col = "red", pch = 20)
    plot(x, yobs, type = "p", pch = 20, xlim = c(0, 200), ylim = c(0.5,
        2), cex = 1)
    lines(z, lwd = 2, col = "red")
    plot(x, y, type = "l", lwd = 2, xlim = c(0, 200), ylim = c(0.5,
        2), col = "blue")
    lines(z, lwd = 2, col = "red")
}
fsimul()
```



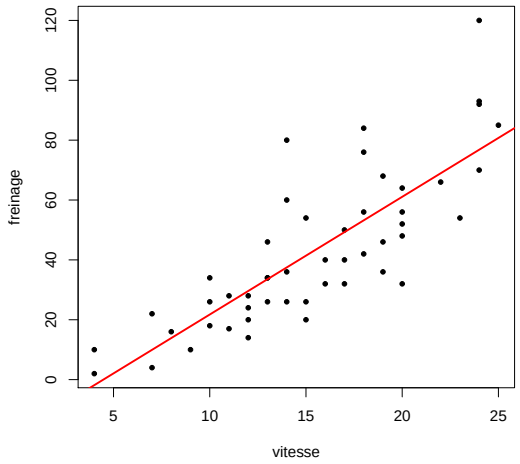

# 3 Quelques exemples et exercices

## 3.1 A propos de sécurité routière

Prenons la relation entre la vitesse des voitures (en miles par heure) et la distance de freinage avant arrêt du véhicule (en pieds). Les données [5] ont été

collectées en 1920 mais restent d'actualité. Superposons le nuage de points et le modèle c'est-à-dire la droite d'ajustement des moindre carrés.

```
data(cars)
plot(cars$speed, cars$dist, xlab = "vitesse", ylab = "distance de\nfreinage",
    pch = 20)
abline(lm(cars$dist ~ cars$speed), lwd = 2, col = "red")
```




Le *coefficient de corrélation linéaire* s'obtient à l'aide de la fonction `cor()` :

```
cor(cars$speed, cars$dist)
```
```
[1] 0.8068949
```

## 3.2 Tension artérielle et fumeurs

Dans une population, on a tiré au sort 34 sujets (17 fumeurs et 17 non fumeurs) à qui on a mesuré la tension artérielle (en mmHg) et demandé l'âge (en années). Les résultats sont dans le tableau 1 et il est temps maintenant d'apprendre à rentrer ses propres données.
Pour cela, deux instructions sont à retenir. Si vous avez déjà un *objet R* (vecteur, facteur, data frame, matrice...), la fonction `edit()` vous permet de visualiser cet objet mais ne vous permet pas de le modifier. Pour cela, il faut utiliser la fonction `fix()`. Si l'objet n'existe pas, créer le en le remplissant de 0 :

```
epidemio <- rep(0, 34 * 3)
epidemio <- data.frame(matrix(epidemio, nrow = 34))
names(epidemio) <- c("tension", "age", "fumeur")
```

Puis utiliser la fonction `fix(epidemio)` pour entrer les valeurs. Vérifier que vous avez bien réalisé toutes les opérations souhaitées :

| | tension | age | fumeur |
|---|---|---|---|
| 1 | 146 | 54 | 1 |
| 2 | 129 | 47 | 1 |
| 3 | 162 | 60 | 1 |
| 4 | 160 | 48 | 1 |
| 5 | 144 | 44 | 1 |
| 6 | 180 | 64 | 1 |
| 7 | 166 | 59 | 1 |
| 8 | 138 | 51 | 1 |
| 9 | 140 | 54 | 1 |
| 10 | 134 | 50 | 1 |
| 11 | 145 | 49 | 1 |
| 12 | 142 | 46 | 1 |
| 13 | 150 | 56 | 1 |
| 14 | 149 | 54 | 1 |
| 15 | 132 | 48 | 1 |
| 16 | 126 | 43 | 1 |
| 17 | 170 | 63 | 1 |
| 18 | 135 | 45 | 0 |
| 19 | 122 | 41 | 0 |
| 20 | 130 | 49 | 0 |
| 21 | 148 | 52 | 0 |
| 22 | 152 | 64 | 0 |
| 23 | 138 | 56 | 0 |
| 24 | 135 | 57 | 0 |
| 25 | 152 | 62 | 0 |
| 26 | 164 | 65 | 0 |
| 27 | 142 | 56 | 0 |
| 28 | 144 | 58 | 0 |
| 29 | 137 | 53 | 0 |
| 30 | 132 | 50 | 0 |
| 31 | 120 | 43 | 0 |
| 32 | 161 | 63 | 0 |
| 33 | 152 | 62 | 0 |
| 34 | 164 | 65 | 0 |

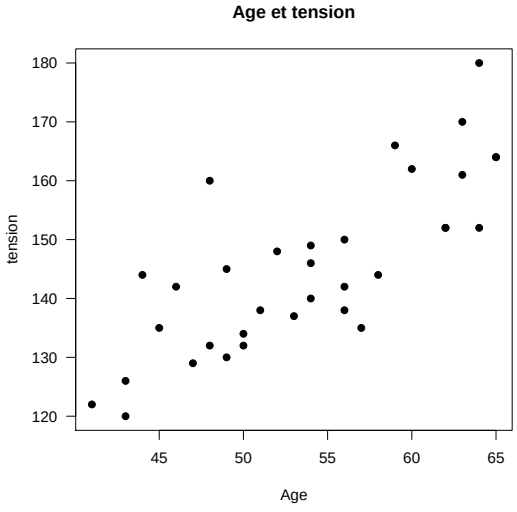
TAB. 1 – Données extraites de Bouyer et al. (1995) Epidémiologie. Principes et méthodes quantitatives, Les éditions INSERM

```
epidemio
```

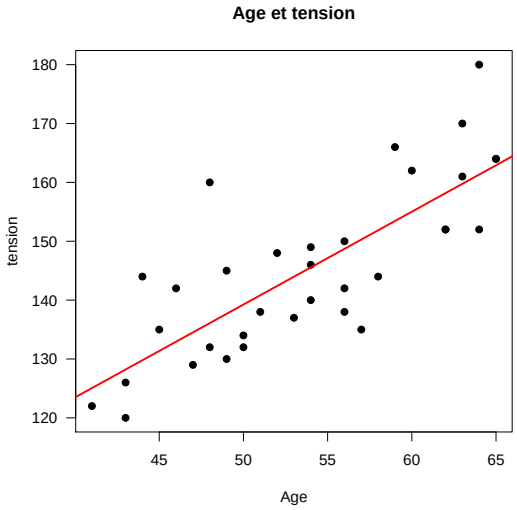
| | tension | age | fumeur |
|---|---|---|---|
| 1 | 146 | 54 | 1 |
| 2 | 129 | 47 | 1 |
| 3 | 162 | 60 | 1 |
| 4 | 160 | 48 | 1 |
| 5 | 144 | 44 | 1 |
| 6 | 180 | 64 | 1 |
| 7 | 166 | 59 | 1 |
| 8 | 138 | 51 | 1 |
| 9 | 140 | 54 | 1 |
| 10 | 134 | 50 | 1 |
| 11 | 145 | 49 | 1 |
| 12 | 142 | 46 | 1 |
| 13 | 150 | 56 | 1 |
| 14 | 149 | 54 | 1 |
| 15 | 132 | 48 | 1 |
| 16 | 126 | 43 | 1 |
| 17 | 170 | 63 | 1 |
| 18 | 135 | 45 | 0 |
| 19 | 122 | 41 | 0 |
| 20 | 130 | 49 | 0 |
| 21 | 148 | 52 | 0 |
| 22 | 152 | 64 | 0 |
| 23 | 138 | 56 | 0 |
| 24 | 135 | 57 | 0 |
| 25 | 152 | 62 | 0 |
| 26 | 164 | 65 | 0 |
| 27 | 142 | 56 | 0 |
| 28 | 144 | 58 | 0 |
| 29 | 137 | 53 | 0 |
| 30 | 132 | 50 | 0 |
| 31 | 120 | 43 | 0 |
| 32 | 161 | 63 | 0 |
| 33 | 152 | 62 | 0 |
| 34 | 164 | 65 | 0 |

**Exercice.**

1. Construire le nuage de points en posant en abscisse l'âge et en ordonnée la tension artérielle.




2. Superposer le modèle




3. Donner ses paramètres

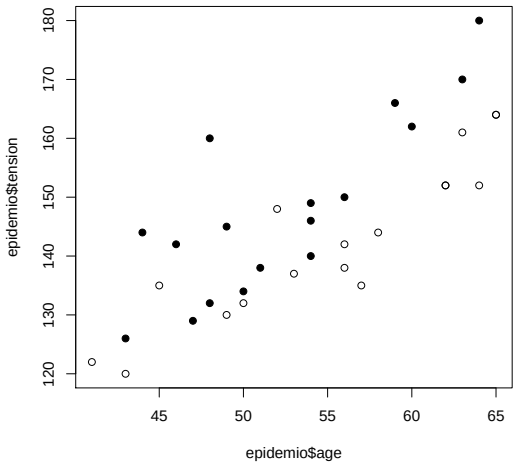
```
(Intercept) epidemio$age
  60.392816     1.577086
```

4. Calculer le coefficient de corrélation linéaire liant ces deux variables.

```
[1] 0.7867896
```

5. Conclure.
6. L'information "fumeur ou non fumeur" n'a pas été introduite. Remplacer le point du nuage par cette information en utilisant les instructions suivantes :

```
plot(epidemio$age, epidemio$tension, pch = c(1, 19)[unclass(factor(epidemio$fumeur))])
```




Les points noirs représentent les fumeurs, les blancs les non fumeurs. Conclure.

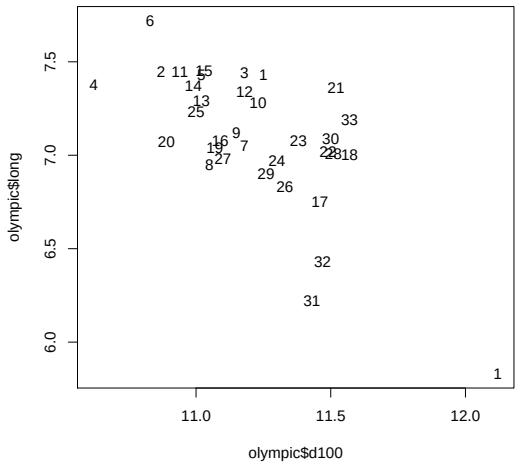
## 3.3 Saut en longueur et 100m

On connait les performances (exemple n° 357 dans [7]) au décathlon masculin de 33 athlètes ayant participé aux Jeux Olympiques de 1988. Les variables sont le 100 m, le saut en longueur, le poids, le saut en hauteur, le 400 m, le 110 m haies, le disque, la perche, le javelot, le 1500 m et score total (utiliser les noms d100, long, poid, haut, d400, d110, disq, perc, jave, d1500 et score).

```
olympic <- read.table("http://pbil.univ-lyon1.fr/R/donnees/olympic.txt")
names(olympic) <- c("d100", "long", "poid", "haut", "d400", "d110",
    "disq", "perc", "jave", "d1500", "score")
names(olympic)
```

```
 [1] "d100"  "long"  "poid"  "haut"  "d400"  "d110"  "disq"  "perc"  "jave"  "d1500"
[11] "score"
```

```
plot(olympic$d100, olympic$long, type = "n")
text(olympic$d100, olympic$long, seq(1:33))
```




**Exercice.**

1. Discuter la représentation graphique.
2. Calculer le coefficient de corrélation. Conclusion.

```
[1] -0.690508
```

3. Discuter de ce résultat et proposer une solution en accord avec les données.

## 3.4 Les données de `Anscombe`

Un des exemples les plus connus de présentation de la relation entre un nuage de points et un coefficient de corrélation concerne les données de Anscombe [2]. Le data.frame contient 8 colonnes, à l'abscisse `x1` correspond l'ordonnée `y1` et ainsi de suite :

```
data(anscombe)
names(anscombe)
```

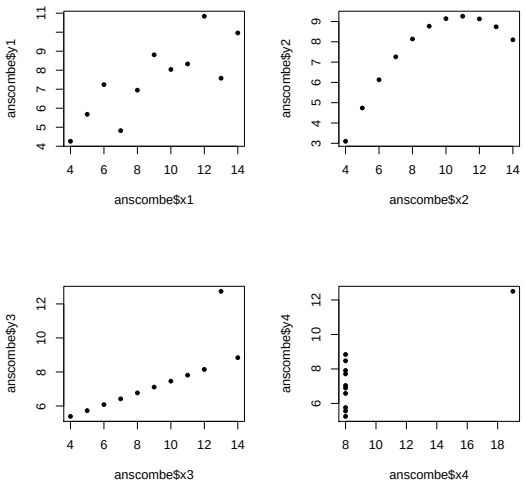
```
[1] "x1" "x2" "x3" "x4" "y1" "y2" "y3" "y4"
```

**Exercice.**

1. Calculer les quatre coefficients de corrélation.

```
[1] 0.8164205 0.8162365 0.8162867 0.8165214
```

2. Décomposer la fenêtre graphique en 4 parties et construire les quatre nuages de points.




3. Commenter.

## 3.5 Mariage et Produit intérieur brut

L'exemple développé concerne le taux de mariages (multiplié par 1000) en fonction du produit intérieur brut [3] de 1974 à 1981.

```
pib <- c(1, 1.1, 1.3, 1.45, 1.7, 1.9, 2.1, 2.4)
taux <- c(7.6, 7.3, 7.2, 6.9, 6.6, 6.3, 6.2, 5.8)
cor(pib, taux)
```

```
[1] -0.9932583
```

**Exercice.** L'auteur soulève la question suivante : "S'agit-il d'un scoop ? l'augmentation du PIB provoquerait-elle une chute de la nuptialité ?". Que pouvez-vous lui répondre ?

## 3.6 Poids du corps et taille du cerveau chez les mammifères

Les données sont extraites de [1].

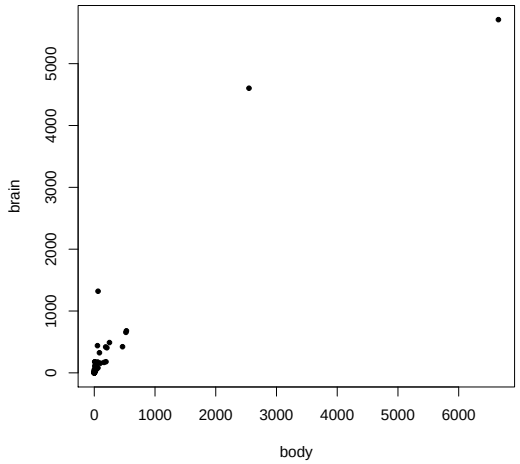
```
library(MASS)
data(mammals)
names(mammals)
```

```
[1] "body"  "brain"
```

```
mammals
```

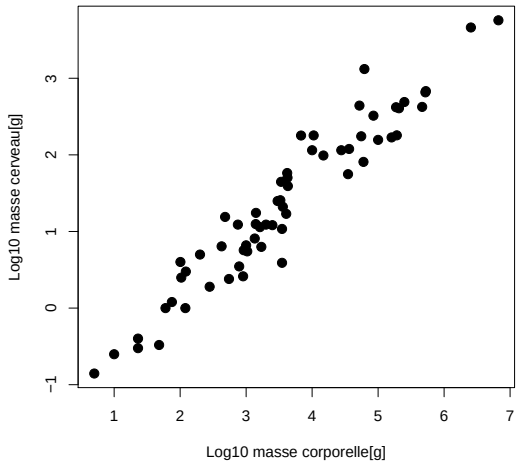
| | body | brain |
|---|---|---|
| Artic fox | 3.385 | 44.50 |
| Owl monkey | 0.480 | 15.50 |
| Mountian beaver | 1.350 | 8.10 |
| Cow | 465.000 | 423.00 |
| Grey wolf | 36.330 | 119.50 |
| Goat | 27.660 | 115.00 |
| Roe deer | 14.830 | 98.20 |
| Guinea pig | 1.040 | 5.50 |
| Verbet | 4.190 | 58.00 |
| Chinchilla | 0.425 | 6.40 |
| Ground squirrel | 0.101 | 4.00 |
| Artic ground squirrel | 0.920 | 5.70 |
| African giant pouched rat | 1.000 | 6.60 |
| Lesser short-tailed shrew | 0.005 | 0.14 |
| Star-nosed mole | 0.060 | 1.00 |
| Nine-banded armadillo | 3.500 | 10.80 |
| Tree hyrax | 2.000 | 12.30 |
| N.A. opossum | 1.700 | 6.30 |
| Asian elephant | 2547.000 | 4603.00 |
| Big brown bat | 0.023 | 0.30 |
| Donkey | 187.100 | 419.00 |
| Horse | 521.000 | 655.00 |
| European hedgehog | 0.785 | 3.50 |
| Patas monkey | 10.000 | 115.00 |
| Cat | 3.300 | 25.60 |
| Galago | 0.200 | 5.00 |
| Genet | 1.410 | 17.50 |
| Giraffe | 529.000 | 680.00 |
| Gorilla | 207.000 | 406.00 |
| Grey seal | 85.000 | 325.00 |
| Rock hyrax-a | 0.750 | 12.30 |
| Human | 62.000 | 1320.00 |
| African elephant | 6654.000 | 5712.00 |
| Water opossum | 3.500 | 3.90 |
| Rhesus monkey | 6.800 | 179.00 |
| Kangaroo | 35.000 | 56.00 |
| Yellow-bellied marmot | 4.050 | 17.00 |
| Golden hamster | 0.120 | 1.00 |
| Mouse | 0.023 | 0.40 |
| Little brown bat | 0.010 | 0.25 |
| Slow loris | 1.400 | 12.50 |
| Okapi | 250.000 | 490.00 |
| Rabbit | 2.500 | 12.10 |
| Sheep | 55.500 | 175.00 |
| Jaguar | 100.000 | 157.00 |
| Chimpanzee | 52.160 | 440.00 |
| Baboon | 10.550 | 179.50 |
| Desert hedgehog | 0.550 | 2.40 |
| Giant armadillo | 60.000 | 81.00 |
| Rock hyrax-b | 3.600 | 21.00 |
| Raccoon | 4.288 | 39.20 |
| Rat | 0.280 | 1.90 |
| E. American mole | 0.075 | 1.20 |
| Mole rat | 0.122 | 3.00 |
| Musk shrew | 0.048 | 0.33 |
| Pig | 192.000 | 180.00 |
| Echidna | 3.000 | 25.00 |
| Brazilian tapir | 160.000 | 169.00 |
| Tenrec | 0.900 | 2.60 |
| Phalanger | 1.620 | 11.40 |
| Tree shrew | 0.104 | 2.50 |
| Red fox | 4.235 | 50.40 |

```
plot(mammals, pch = 20)
```




Assurément, les espèces ayant un gros corps ont tendance à avoir un gros cerveau (encore un scoop !). Mais comment pourrions nous exprimer cette relation ? Nous allons convertir les données initiales en logarithmes de base 10 avant de les représenter, comme dans la figure ci-après (NB : il est plus commode d'avoir les mêmes unités, la masse corporelle sera maintenant exprimée en grammes).

```
x <- log10(1000 * mammals$body)
y <- log10(mammals$brain)
plot(x, y, pch = 20, cex = 2, xlab = "Log10 masse corporelle[g]",
    ylab = "Log10 masse cerveau[g]")
```




Ce dernier exemple est extrait d'une fiche sur l'allométrie que vous trouvez à l'adresse suivante : `http://pbil/R/fichestd/tdr333.pdf`.

# 4 Conclusion

L'existence d'une corrélation élevée entre deux variables $x$ et $y$ ne conduit pas à l'existence d'une relation **cause - effet**. On utilise la connaissance de $x$ pour prédire des valeurs de $y$. Cela n'implique pas qu'un changement de $x$ cause un changement de $y$. Considérons par exemple le dicton : "Une pomme par jour garde le médecin éloigné". Une corrélation négative, modérée peut être trouvée sans aucun doute entre le nombre de pommes mangées et le nombre de visites chez le mdédecin. Cela n'implique pas qu'une personne va fréquemment chez le médecin parce qu'elle mange un nombre insuffisant de pommes.

Considérons un autre exemple du genre. Dans " Une logique de la communication ", Paul Watzlawick `http://www.evoweb.net/stat.htm` raconte que la plus forte corrélation trouvée dans les années 1950 a été celle entre la consommation de bière sur la côte ouest des USA, et la mortalité infantile au Japon. Cet exemple a été fréquemment repris pour montrer les limites des statistiques et démontrer " qu'on peut leur faire dire n'importe quoi ". Et en effet beaucoup feront remarquer qu'on ne peut accuser les Américains assoiffés de tuer les Japonais (on remarquera d'ailleurs que personne n'accuse les enfants Japonais d'assoiffer les Américains).

> Ne jamais confondre co-relation et relation cause - effet. Le coefficient de corrélation indique l'existence et la nature d'une relation entre deux variables. L'interprétation ne peut se faire que dans le contexte dans lequel les variables sont analysées.

# Références

[1] T. Allison and D. V. Cicchetti. Sleep in mammals : ecological and constitutional correlates. *Science*, 194 :732–734, 1976.

[2] F. J. Anscombe. Graphs in statistical analysis. *American Statistician*, 27 :17–21, 1973.

[3] Robert C. *Contes et décomptes de la statistique. Une initiation par l'exemple.* Vuibert, Paris, France, 2003.

[4] J.J. Droesbeke and Ph. Tassi. *Histoire de la Statistique.* Que sais-je ? P.U.F., Paris, 1990.

[5] M. Ezekiel. *Methods of Correlation Analysis.* Wiley, New York, USA, 1930.

[6] F. Galton. Regression towards mediocrity in hereditary stature. *Journal of Anthropological Institute*, 15 :246–263, 1886.

[7] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski. *A handbook of small data sets.* Chapman & Hall, London, 1994.

[8] K. Pearson. Contributions to the mathematical theory of evolution. iii regression, heredity and, panmixia. *Philosophical Transactions of the Royal Society London Series A*, 187 :253–318, 1896.