Fiche TD avec le logiciel R : bem3

De la stature chez l'Homme... à la taille des cerveaux chez les mammifères. Réversion, régression, corrélation

A.B. Dufour, Pr Jean R. Lobry, D. Chessel, N. Rochette

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).

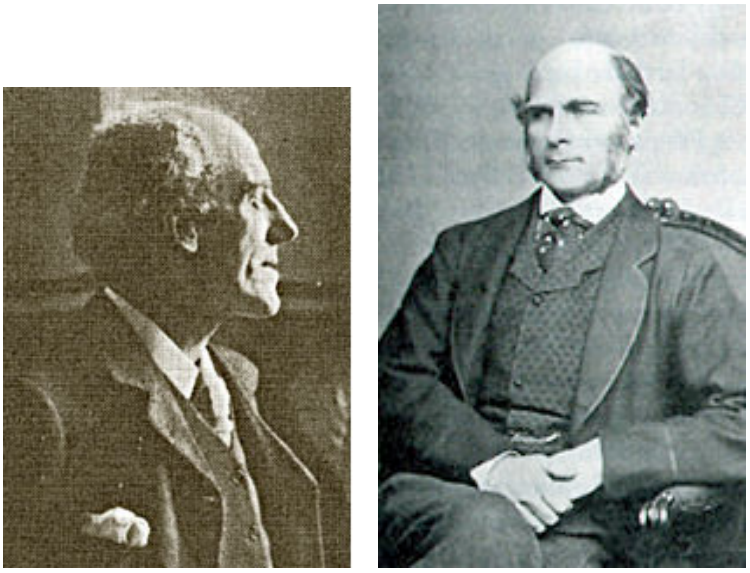


FIGURE 1 – K. Pearson et F. Galton

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.*

En français cela donne : « La réversion est une tendance de la part du type filial moyen à s'éloigner du type parental, et à retourner à ce qui pourrait être décrit comme l'état ancestral moyen. Si la variabilité familiale avait été le seul processus à affecter les charactéristiques d'un échantillon au fil du temps, la dispersion de la race autour de sa moyenne idéale croîtrait indéfiniment avec le nombre de générations. Mais la réversion frein cette accroissement, et lamène à un point stable. »

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.*

En français encore : « Il est facile de voir que cette co-relation doit être la conséquence des variations de deux organes, partiellement due aux mêmes causes. Si ces variations étaient intégralement dues aux mêmes causes, la co-relation devrait être parfaite, comme c'est approximativement le cas avec la symétrie globale du corps. S'il n'y avait aucune cause commune, la co-relation devrait être nulle. Entre ces deux extrêmes on trouve un nombre infini de cas intermédiaires et on va montrer que l'intensité de la co-relation dans chaque cas peut être exprimée par une unique valeur. »

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.

2 La corrélation

2.1 Qu'est-ce que c'est ?

On dit que deux variables X et Y sont *corrélées* quand il existe un lien entre leurs valeurs. Par exemple, si les valeurs de X et Y tendent à être toujours grandes ensemble, ou petites ensemble, ou encore si celles d'X sont toujours petites quand celles d'Y sont grandes, alors X et Y sont corrélées. Les couples de variables $Y = 2X$ et $Y = -X$ sont corrélés, et on parle de corrélation positive dans le premier cas et négative dans le second.

Plus généralement, une corrélation linéaire entre deux variables signifie qu'il est possible de prédire –partiellement au moins– une variable par l'autre à l'aide d'un modèle linéaire, c'est à dire de type $Y = aX + b$.

En statistiques, on mesure la corrélation linéaire par le *coefficient de corrélation* de Pearson :

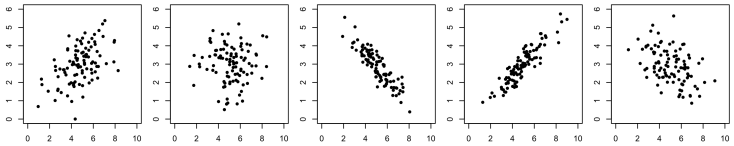
$$r = \frac{cov(X,Y)}{sd(X) \cdot sd(Y)} \left(= \frac{\sum (x - \bar{x}) \cdot (y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \cdot \sqrt{\sum (y - \bar{y})^2}} \right)$$

Dans la formule précédente, *cov* signifie covariance entre X et Y. Le coefficient de corrélation varie entre -1 et $+1$. Il est nul quand les variables sont indépendantes, négatif quand les variables sont corrélées négativement (ie. quand la

valeur de X est grande, celle d'Y est petite) et positif quand elles sont corrélées positivement.

Pour s'affranchir du signe de la corrélation, on peut aussi utiliser la mesure r^2 ("r carré") qui varie entre 0 (pas de corrélation linéaire) et 1 (corrélation parfaite).

Exercice. Qui est qui ? Retrouvez quel graphique va avec quel coefficient de corrélation.



```
[1] -0.89 -0.49 -0.01  0.50  0.91
```

Il est important de remarquer que, si le coefficient de corrélation linéaire de deux variables indépendantes est toujours nul, *la réciproque est fausse* : un coefficient de corrélation linéaire nul n'implique pas que les variables sont indépendantes. L'exemple classique est $Y = X^2$; les deux variables ne sont clairement pas indépendantes pourtant leur coefficient de corrélation *linéaire* est nul.

Exercice

Examinons les données [6] de Francis Galton (1822-1911) sur la relation entre la taille (en pouces) 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.

Vous trouverez ces données à l'adresse suivante : `http://pbil.univ-lyon1.fr/R/donnees/taimdc.txt`. Importez-les dans R et enregistrez-les dans un objet appelé `taimdc`. Elles sont en pouces, ce n'est pas très lisible pour nous français, aussi convertissez-les en centimètres (1 pouce vaut 2.54 cm).

Vos données semblent-elles correctes ?

```
summary(taimdc)
```

Calculez le coefficient de corrélation entre la taille du mid-parent et celle des enfants. D'après vous, existe-t-il une corrélation entre ces deux variables ?

La fonction `cor()` calcule le coefficient de corrélation entre deux variables.

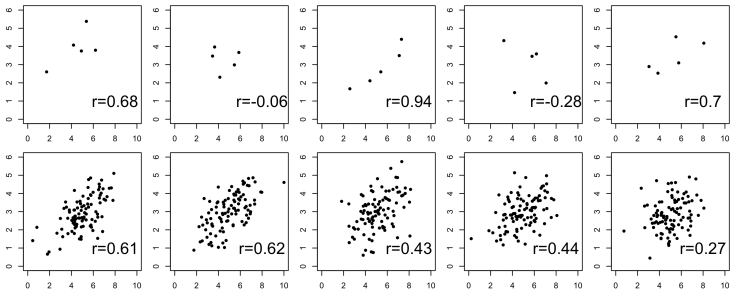
2.2 Erreur statistique

En statistique, on ne connait jamais exactement la distribution exacte d'une variable, et il nous est nécessaire de l'approcher à partir d'un *échantillon*. D'une manière générale, plus l'échantillon est grand, plus les conclusions seront précises. Si l'échantillon est trop petit, on ne peut rien conclure, et c'est vrai en particulier pour la corrélation. Le problème est : peut-on donner une valeur

fiable de la corrélation entre deux variables si on ne dispose pas d'assez de points concernant ces deux variables ?

Si l'on a que quatre ou cinq points, le calcul du coefficient du corrélation est très imprécis parce qu'avec seulement cinq points, les variables sont très mal décrites.

Exemple : On considère deux variables X et Y ayant un coefficient de corrélation de $+0.5$. Pour les graphiques ci-dessous, on a simulé à chaque fois cinq réalisations de ces deux variables (*ie.* cinq points), puis 100 réalisations, et calculé le coeffeicient de corrélation estimé :



Ces simulations montrent que lorsqu'on a que cinq points, l'erreur statistique est très grande. R implémente les outils statistiques qui permettent de quantifier l'imprécison sur le coefficient de corrélation dans la fonction `cor.test()`. Reprenons le jeu de données de Galton sur la corrélation entre les tailles des parents et des enfants :

```
cor.test(taimdc$x, taimdc$y)
```

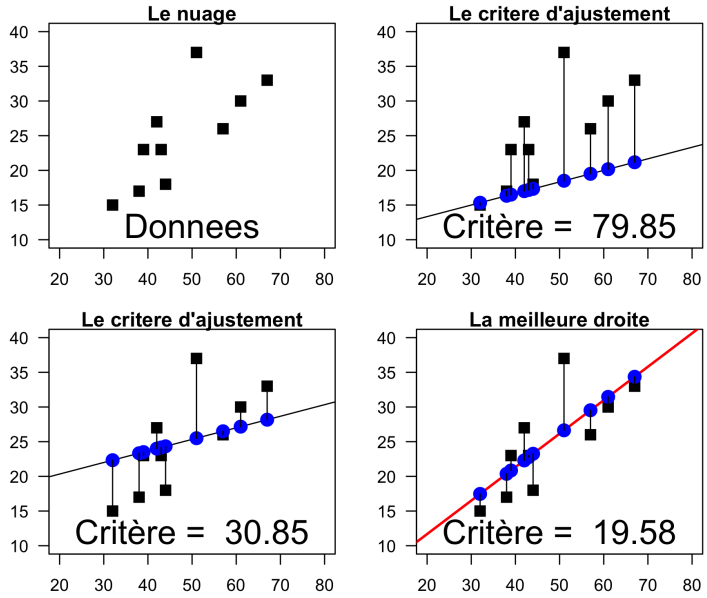
```
        Pearson's product-moment correlation
data:  taimdc$x and taimdc$y
t = 20, df = 900, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.406 0.508
sample estimates:
  cor
0.459
```

Quel est l'*intervalle de confiance* ('confidence interval' en anglais) de la valeur du coefficient de corrélation ? Qu'en déduisez-vous quant à l'existence d'une corrélation entre les deux variables ?

3 Régression

Lorsque deux variables sont corrélées, il est utile de faire une **régression linéaire** : on met dans le nuage une droite qui s'ajuste au mieux. Cela revient à déterminer le modèle linéaire optimal pour prédire Y avec X. Le critère est celui des moindre carrés : le modèle 'optimal' est celui qui minimise la moyenne des carrés des **résidus** – les résidus sont les écarts entre les valeurs observées et celles prédites par le modèle linéaire, ils sont donc représentés par les barres verticales sur les graphiques ci-dessous.

La droite ainsi obtenue s'appelle **droite de régression**.



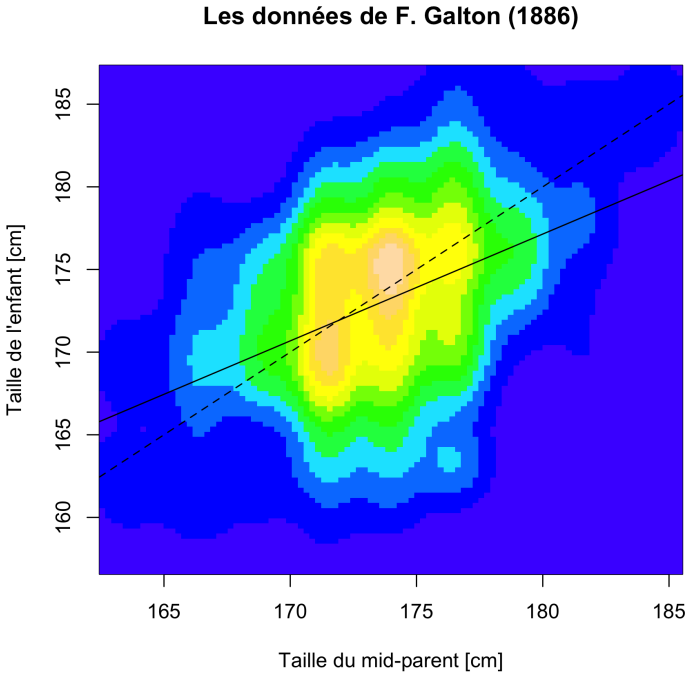
Exercice

Représentez graphiquement les données de Galton (`taimdc`). Quelle fonction graphique avez-vous utilisée ? Est-ce que le graphique obtenu est cohérent avec la taille du jeu de données ?

Les points étant nombreux, on peut utiliser `sunflowerplot()` ou une carte de densité – on utilisera ici `kde2d()` pour le calcul de la densité et `image()` pour sa représentation. On se propose d'ajouter sur le graphique la droite d'équation $Y = X$ (les enfants font la même taille que leurs parents) et la droite de régression calculée sur les données.

Les régressions linéaires se font avec la fonction `lm()`, pour *linear model*.

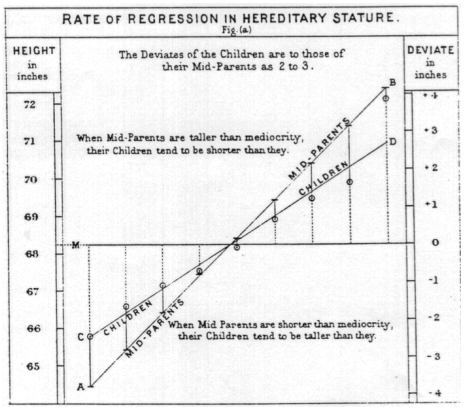
```
# On calcule et on représente la densité de points
densite2d <- kde2d(x = taimdc$x, y = taimdc$y, n = 100)
image(densite2d,
  col = topo.colors(12),
  xlab = "Taille du mid-parent [cm]",
  ylab = "Taille de l'enfant [cm]",
  main = "Les données de F. Galton (1886)"
)
# La droite d'équation Y=X
abline(c(0,1), lty="dashed") # (lty:"line type")
# Régression linéaire et tracé de la droite de régression
modele_lin <- lm(taimdc$y ~ taimdc$x)
abline(modele_lin)
```



Les coefficients de la droite de régression sont donnés par :

```
coefficients(modele_lin)
```
```
(Intercept)    taimdc$x
     60.811       0.646
```

On constate que la pente de la droite (0.65) est inférieure à 1. 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. De manière plus générale, on parlera de corrélation quand on est interessé par l'*existence* d'un lien entre deux variables, et de régression quand c'est l'intensité de ce lien qui nous précoccupe le plus. Mais les deux termes se recouvrent.

4 Quelques exemples et exercices

4.1 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. À l'aide de la fonction `data()`, chargez le jeu de données `cars`, et lisez la documentation de ce jeu de données. Convertissez les vitesses en km/h et les distances en mètres :

```
data(cars)
head(cars)
```

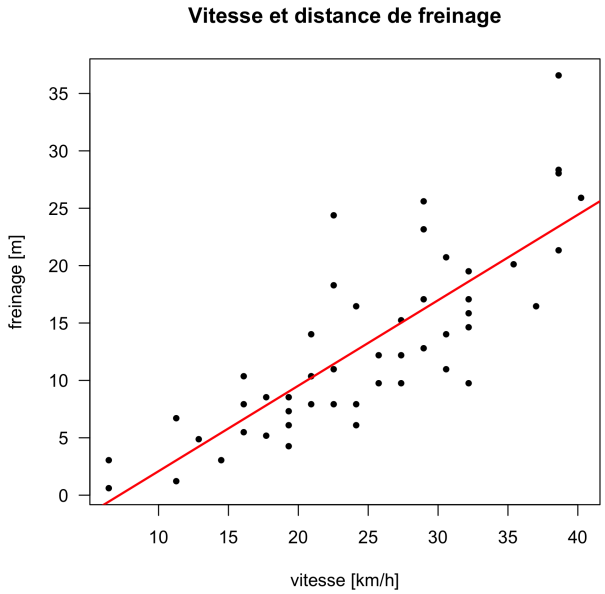
```
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

Après conversion, vous devriez obtenir quelque chose qui ressemble à ca :

```
head(cars)
```

```
  speed dist vitesse distance
1     4    2    6.44     0.61
2     4   10    6.44     3.05
3     7    4   11.27     1.22
4     7   22   11.27     6.71
5     8   16   12.87     4.88
6     9   10   14.48     3.05
```

Tracez le graphique représentant la distance de freinage en fonction de la vitesse, en mettant des titres aux axes. Comme vous avez fait pour les données de Galton, superposez le modèle linéaire correspondant.



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

```
cor(cars$speed,cars$dist)
```

```
[1] 0.807
```

Qu'en déduisez vous quant à la force de la relation linéaire entre les deux variables ?

4.2 Tension artérielle et fumeurs

	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

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

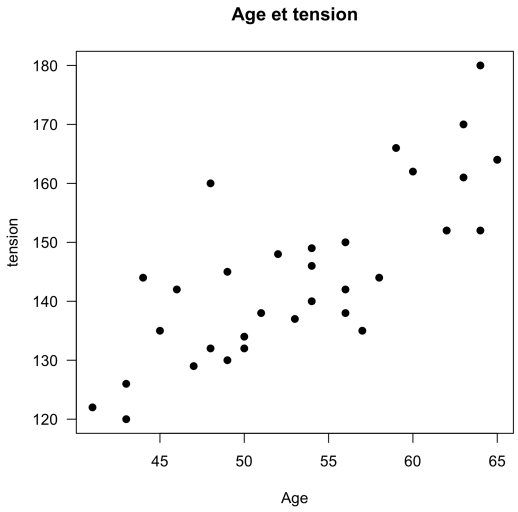
Dans une population, on a tiré au sort 34 sujets (17 fumeurs et 17 non fumeurs) à qui on a mesuré la tension artérielle (en mmHg) et demandé l'âge (en années). Les résultats sont dans le tableau 1 et il est temps maintenant d'apprendre à rentrer ses propres données.
Pour cela, une manière de faire consiste à rentrer les données du tableau 1 dans un fichier texte (prenez l'exemple de `t3var.txt`) et ensuite à les charger dans R avec la commande `read.table` sous le nom de `epidemio`.

Vérifiez que l'objet `epidemio` est bien créé :

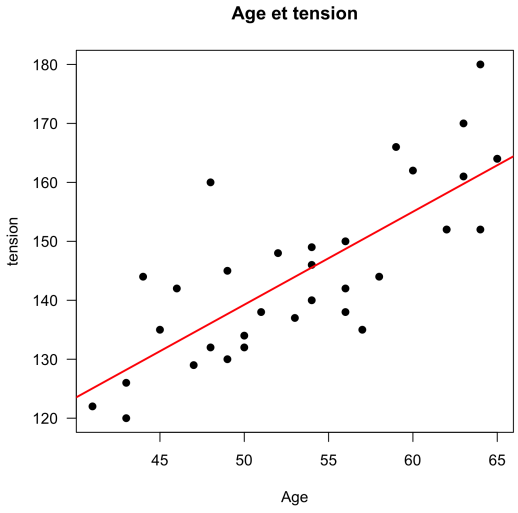
```
epidemio
   tension age fumeur
1      146  54      1
2      129  47      1
3      162  60      1
4      160  48      1
5      144  44      1
6      180  64      1
7      166  59      1
8      138  51      1
9      140  54      1
10     134  50      1
11     145  49      1
12     142  46      1
13     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
      60.39         1.58
```

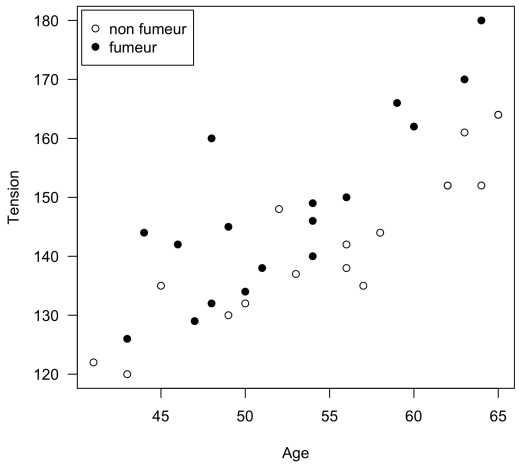
4. Calculer le coefficient de corrélation linéaire liant ces deux variables.

```
[1] 0.787
```

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 :

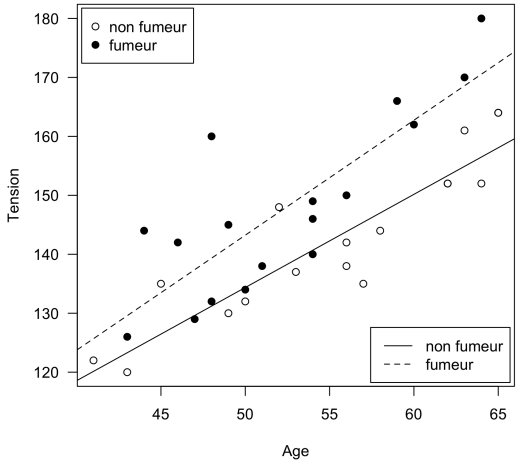
```
plot(epidemio$age[epidemio$fumeur==0],
epidemio$tension[epidemio$fumeur==0],pch=1,
xlim = range(epidemio$age), ylim = range(epidemio$tension), las = 1,
```

```
xlab = "Age", ylab = "Tension")
points(epidemio$age[epidemio$fumeur==1],epidemio$tension[epidemio$fumeur==1],pch=19)
legend("toplef",inset=0.01, c("non fumeur","fumeur"), pch = c(1,19))
```



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

7. Reprendre le graphique précédent et y ajouter les droites de régression séparément pour les fumeurs et non fumeurs :



Quel est l'intérêt des droites de régression ici ?

4.3 Saut en longueur et 100m

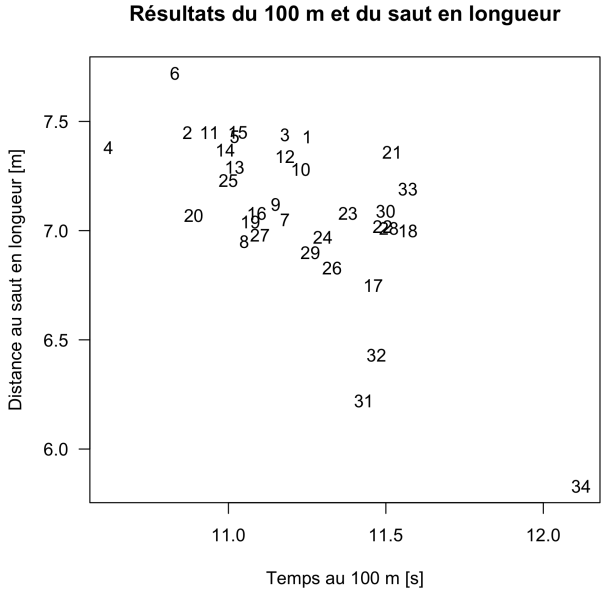
On connait les performances (exemple n° 357 dans [7]) au décathlon masculin de 34 athlètes ayant participé aux Jeux Olympiques de 1988. Les variables sont le 100 m, le saut en longueur, le poids, le saut en hauteur, le 400 m, le 110 m

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")
head(olympic)
```

```
  d100 long poid haut d400 d110 disq perc jave d1500 score
1 11.2 7.43 15.5 2.27 48.9 15.1 49.3  4.7 61.3   269  8488
2 10.9 7.45 15.0 1.97 47.7 14.5 44.4  5.1 61.8   273  8399
3 11.2 7.44 14.2 1.97 48.3 14.8 43.7  5.2 64.2   263  8328
4 10.6 7.38 15.0 2.03 49.1 14.7 44.8  4.9 64.0   285  8306
5 11.0 7.43 12.9 1.97 47.4 14.4 41.2  5.2 57.5   257  8286
6 10.8 7.72 13.6 2.12 48.3 14.2 43.1  4.9 52.2   274  8272
```

```
plot(olympic$d100, olympic$long, type="n", xlab = "Temps au 100 m [s]",
ylab = "Distance au saut en longueur [m]", las = 1,
main = "Résultats du 100 m et du saut en longueur")
text(x = olympic$d100, y = olympic$long, 1:nrow(olympic))
```



Exercice.

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

```
[1] -0.691
```

3. D'après cette représentation graphique, quels sont les meilleurs sportifs et les moins bons ?

4.4 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 :

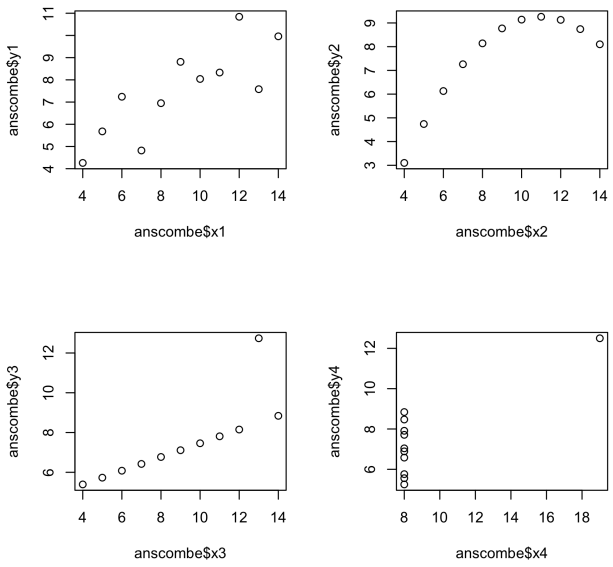
```
data(anscombe)
names(anscombe)
```
```
[1] "x1" "x2" "x3" "x4" "y1" "y2" "y3" "y4"
```

Exercice.

1. Calculer les quatre coefficients de corrélation.

   ```
   [1] 0.816 0.816 0.816 0.817
   ```

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



3. Commenter.

4.5 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] -0.993
```

Exercice. L'auteur soulève la question suivante : « S'agit-il d'un scoop ? l'augmentation du PIB provoquerait-elle une diminution du nombre de mariages ? ». Que pouvez-vous lui répondre ?

4.6 Piraterie et réchauffement climatique

Un des dogmes de la religion pastafariste (si, si) est que le réchauffement climatique est une conséquence directe de la diminution du nombre de pirates. Cette assertion est prouvée par les données suivantes :

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

```
  an   ndp temp
1 1820 35000 14.2
2 1860 45000 14.3
3 1880 20000 14.6
4 1920 15000 14.9
5 1940  5000 15.2
6 1980   400 15.6
7 2000    17 15.9
```

```
cor(pirates$ndp, pirates$temp)
```

```
[1] -0.926
```

Étudiez ce petit jeu de données (les trois colonnes sont le nombre de pirates, la température mondiale moyenne et l'année), et discutez de cette affirmation d'un point de vue scientifique, (et non religieux).

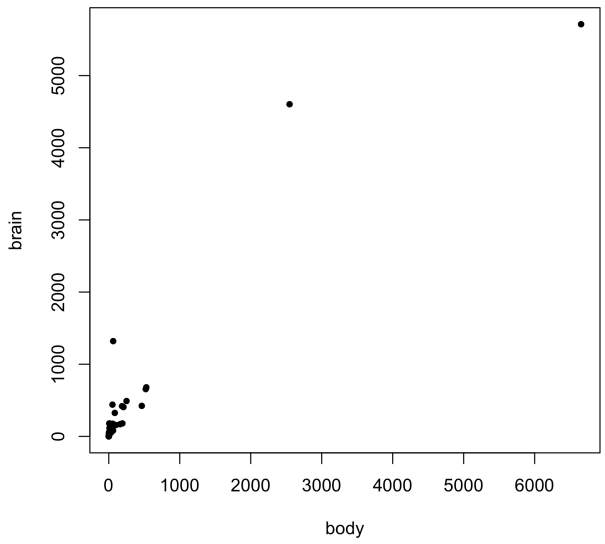
4.7 Poids du corps et taille du cerveau chez les mammifères

Les données sont extraites de [1].

```
library(MASS)
data(mammals)
head(mammals)
```

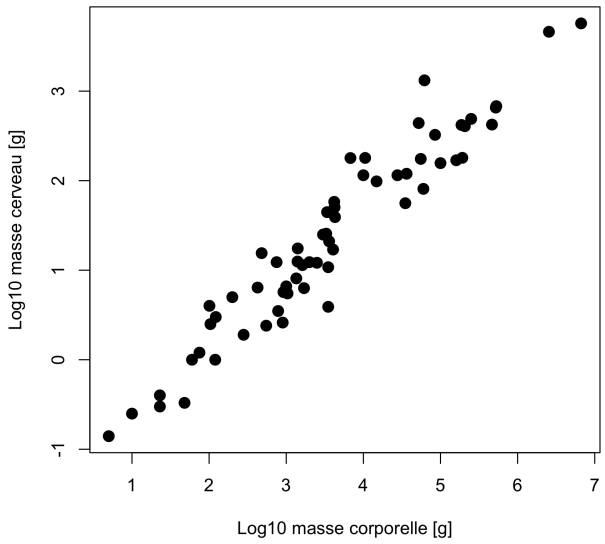
```
                body brain
Arctic fox      3.38  44.5
Owl monkey      0.48  15.5
Mountain beaver 1.35   8.1
Cow           465.00 423.0
Grey wolf      36.33 119.5
Goat           27.66 115.0
```

```
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.

Transformez la masse corporelle en grammes. Ensuite, convertissez les deux variables en logarithme de base 10. Tracez le graphique qui représente le log10 de la masse du cerveau en fonction du log10 de la masse corporelle.



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

5 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, editors. *A handbook of small data sets.* Chapman & Hall, London, 1984.

[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.