L3 Pro "Biotechnologies végétales et création variétale"
TP 4 : Corrélation et Régression

A-B. Dufour, J.R. Lobry, D. Chessel, N. Rochette, M. Bailly-Bechet

Automne 2012

1 La corrélation

1.1 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 $Y = aX + b$.

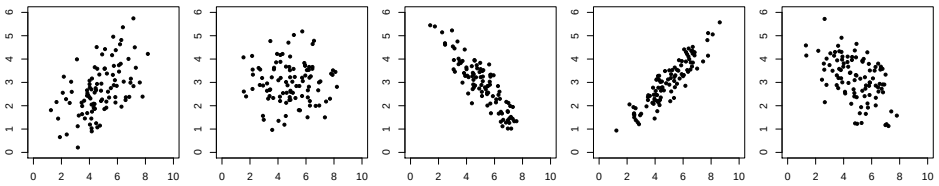
En statistiques, on mesure la corrélation linéaire par le *coefficient de corrélation* de Pearson :

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

Ce coefficient 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] -0.90 -0.49  0.00  0.52  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 de Francis Galton (1822-1911), un des pionniers de l'analyse des corrélations, sur la relation entre la taille (en pouces) de 928 enfants et la taille de leurs parents (en pouces). Dans le jeu de données, la première colonne contient la taille moyenne des parents [1], dit mid-parent, la seconde colonne celle des enfants.

Vous trouverez ces données à l'adresse suivante : `http://pbil.univ-lyon1.fr/members/mbailly/TP/taimdc.xls`. Ces données sont dans un fichier Excel, que R ne sait pas lire directement. Il faut les convertir à un format lisible. Pour cela :

- Sauvegardez le fichier Excel dans votre répertoire, et ouvrez-le avec OpenOffice.
- Dans "Enregistrer sous", choisissez dans la rubrique "Filtre" l'option "Texte CSV"
- OpenOffice vous demande comment représenter les changements de colonne, avec l'option "Séparateur de champ". Choisissez "Tabulation".
- Sauvegardez. Vous avez un fichier avec l'extension `.csv` qui est un fichier au format texte, donc lisible par R ainsi que par n'importe quel autre ordinateur ou logiciel d'analyse.

Pour importer ce fichier dans R, une petite précision : vous avez vu que les nombres comportent des virgules, et pas des points. Il faut le préciser à R lors de la lecture avec l'option `dec`, pour qu'il reconnaisse des nombres ; la commande de lecture donne donc :

```
taimdc<-read.table("taimdc.csv",header=TRUE,dec=",")
```

```
head(taimdc)
```

```
     x    y
1 73.0 72.2
2 73.0 73.2
3 73.0 73.2
4 73.0 73.2
5 72.5 68.2
6 72.5 69.2
```

Notez comme R a converti les nombres à virgule en nombres avec un point, à l'anglaise. Autre question nationale, les tailles sont en pouces, ce n'est pas très lisible pour nous français, aussi convertissez-les en centimètres (1 pouce vaut 2.54 cm).

1. Galton a effectué une correction pour ne pas être gené par les variations dues au sexe de l'enfant, qui est déja incluse dans ce jeu de données avec la taille "moyenne" des parents

Vos données semblent-elles correctes ?

```
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  
```

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 ?

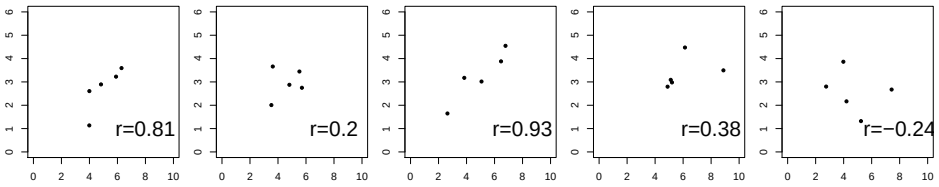
La fonction `cor()` calcule le coefficient de corrélation entre deux variables.

1.2 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.

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 réel de $+0.5$. Pour les graphiques ci-dessous, on a simulé à chaque fois cinq réalisations de ces deux variables (*ie.* cinq points) :



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()`. Cette fonction effectue un test du coefficient de corrélation. Les hypothèses de ce test sont :

H_0 : les variables X et Y *ne sont pas* linéairement reliées (i.e. elles sont *indépendantes*) et le coefficient de corrélation r *n'est pas* significativement différent de 0.

H_1 : les variables X et Y *sont* linéairement reliées (i.e. elles sont *dépendantes*) et le coefficient de corrélation r *est* significativement différent de 0.

Reprenons le jeu de données de Galton sur la corrélation entre les tailles des parents et des enfants :

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

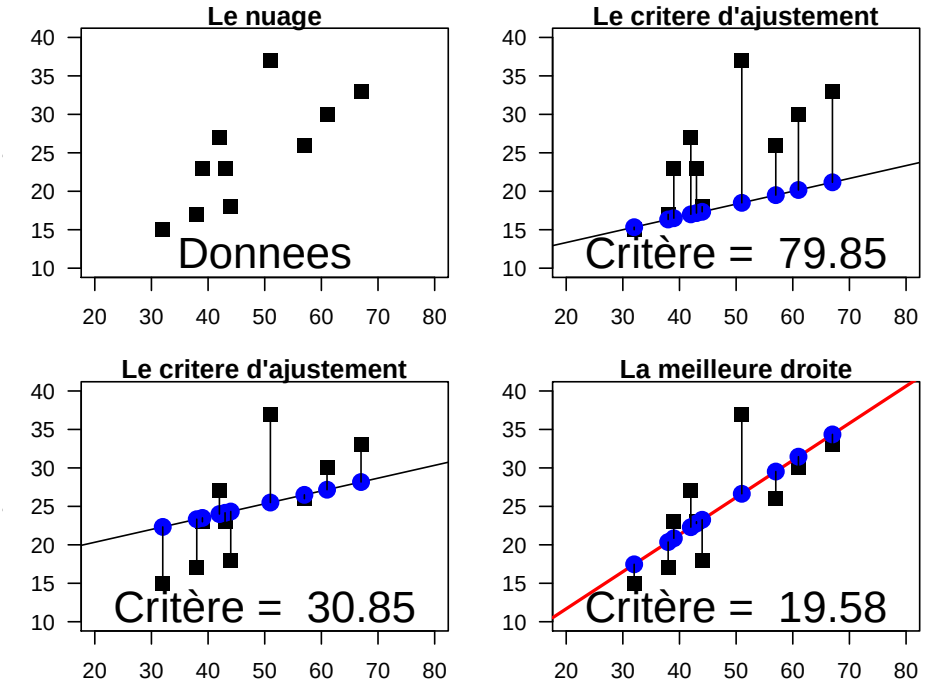
```
        Pearson's product-moment correlation

data:  taimdc$x and taimdc$y
t = 15.7111, df = 926, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4064067 0.5081153
sample estimates:
      cor
0.4587624
```

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

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



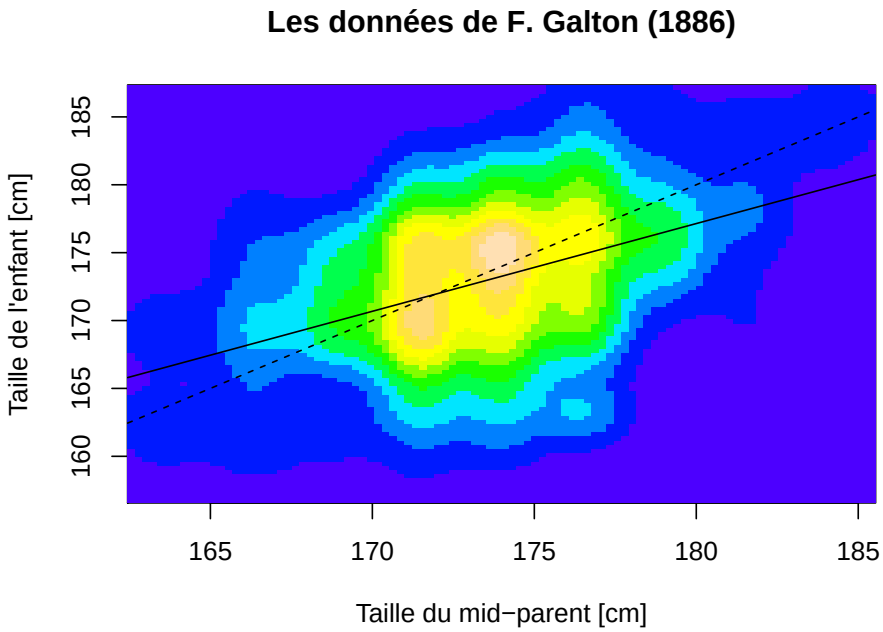
Exercice

Représentez graphiquement les données de Galton (`taimdc`). Quelle difficulté observe-t-on en faisant le graphique naïvement ?

Les points étant nombreux, on peut utiliser 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*.

```
# Les lignes commencant par un # sont des commentaires
# Les taper ne sert à rien
# 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")
# 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
 60.8114867   0.6462906
```

On constate que la pente de la droite (0.646) 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.

2.1 Conditions d'application

Un point est à garder en mémoire quand on effectue une régression linéaire : pour qu'une droite soit un "bon" modèle pour la relation entre Y et X, il faut que les points sur le graphe soient raisonnablement alignés. Si on calcule le coeffcient de corrélation entre X et $Y = e^X$, il sera positif, significativement différent de 0, mais une droite ne sera pas un bon modèle pour la relation observée !

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. Le `data.frame` est pré-existant dans R, contient 8 colonnes, à l'abscisse `x1` correspond l'ordonnée `y1` et ainsi de suite :

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

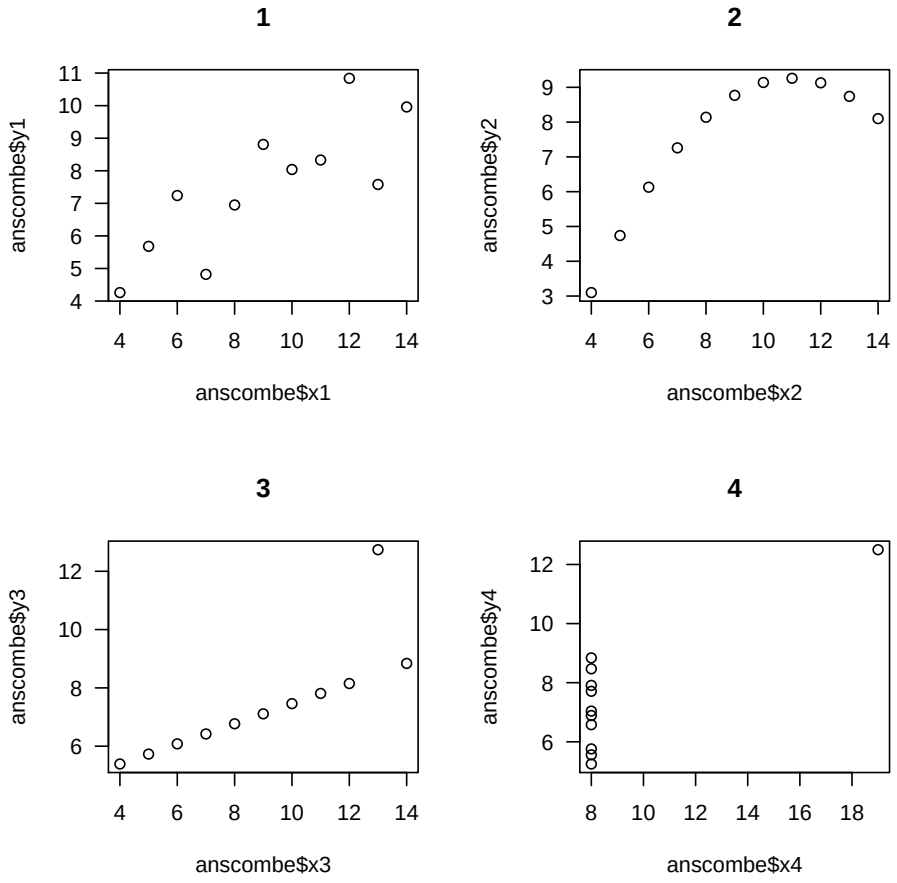
```
[1] "x1" "x2" "x3" "x4" "y1" "y2" "y3" "y4"
```

Exercice.

1. Calculer les quatre coefficients de corrélation. Que remarquez-vous ?

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

   La représentation graphique de ces 4 jeux de données est la suivante :



2. Comparez la variabilité de ces représentations graphiques en relation avec la proximité de leurs coefficients de corrélation. Que pouvez-vous conclure ? Identifiez ce qui, dans chaque graphique, pourrait violer les conditions d'application de l'analyse par régression linéaire.

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 ont été collectées en 1920 mais restent d'actualité. À l'aide de la fonction `data()`, chargez le jeu de données `cars`, puis convertissez les vitesses en km/h et les distances en mètres :

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

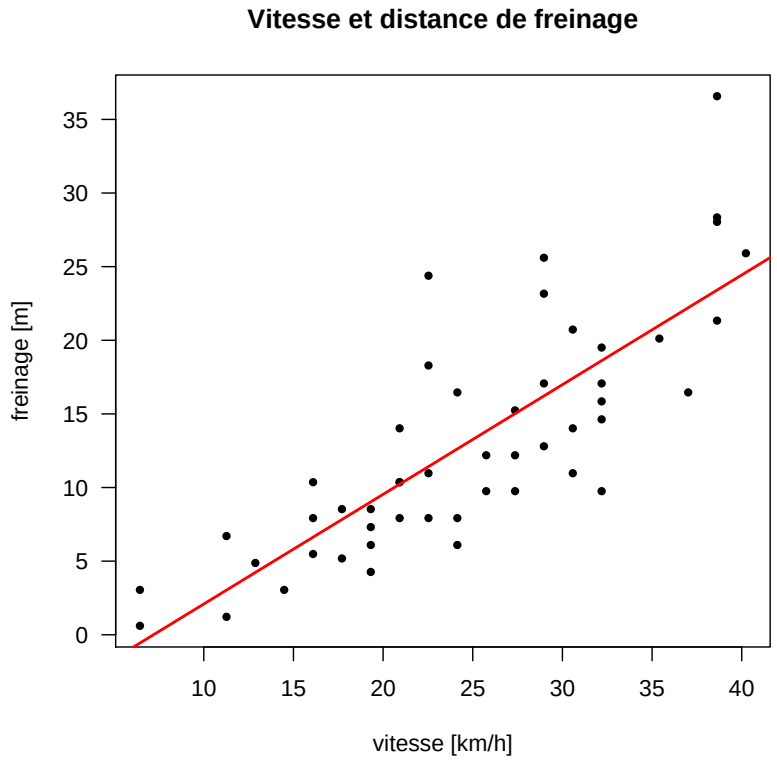
```
  speed dist
1     4    2
```

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

```
cars$vitesse <- cars$speed*1.609344
cars$distance <- cars$dist*0.3048
head(cars)
```

```
  speed dist   vitesse distance
1     4    2  6.437376   0.6096
2     4   10  6.437376   3.0480
3     7    4 11.265408   1.2192
4     7   22 11.265408   6.7056
5     8   16 12.874752   4.8768
6     9   10 14.484096   3.0480
```

Tracez le graphique représentant la distance de freinage en fonction de la vitesse, en mettant des titres aux axes. Vous devriez obtenir des points assez bien alignés pour que l'on puisse vouloir tracer le modèle linéaire correspondant. Comme vous avez fait pour les données de Galton, superposez le modèle linéaire correspondant au graphe.



Calculez le coefficient de corrélation linéaire. Qu'en déduisez vous quant à la force de la relation linéaire entre les deux variables ?

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 à l'adresse `http://pbil.univ-lyon1.fr/members/mbailly/TP/epidemio.xls`. Faites ce qu'il faut pour charger ces données dans une variable du nom de `epidemio`.

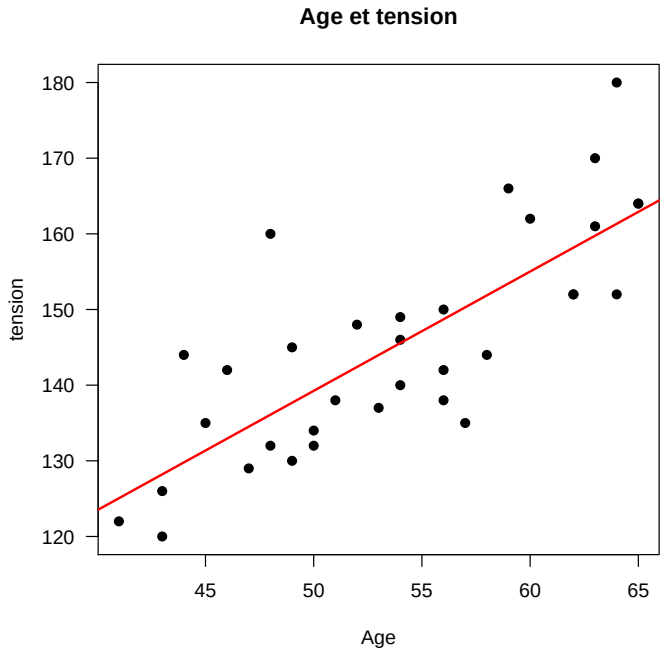
Vérifiez que l'objet `epidemio` est bien créé :

```
head(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
```

Exercice.

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



2. Donner ses paramètres

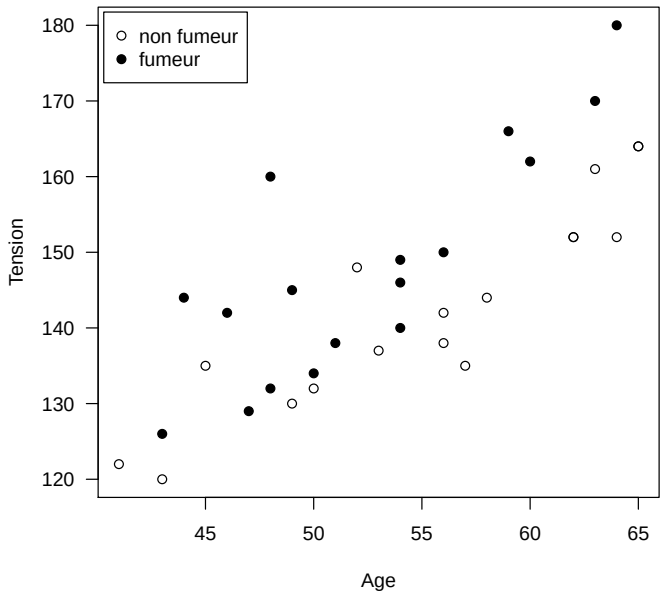
```
(Intercept) epidemio$age
  60.392816     1.577086
```

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

```
[1] 0.7867896
```

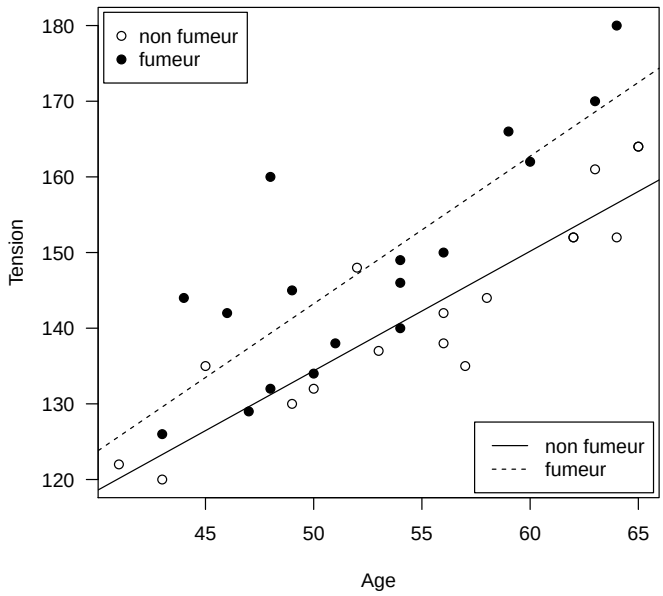
4. Conclure.
5. L'information "fumeur ou non fumeur" n'a pas été introduite. Coloriez les points du nuage par cette information en utilisant les instructions suivantes :

```
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("topleft",inset=0.01, c("non fumeur","fumeur"), pch = c(1,19))
```



Les points noirs représentent les fumeurs, les blancs les non fumeurs. Sur la base de ce graphique, que peut-on conclure ?

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



Comment interpréter les pentes et les ordonnées à l'origine de ces deux droites de régression ici ? Biologiquement, que pouvez-vous conclure à partir du graphe précédent ?

3.3 Piraterie et réchauffement climatique

Un des dogmes de la religion (parodique) pastafariste 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] -0.9261274
```

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

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 un exemple classique 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). Ici les causes de la variation commune de ces deux variables sont à chercher dans le contexte historique de l'après-guerre...

Ne jamais confondre corrélation 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.