Chapitre V : Coefficients de corrélation et tests

Cours de tests paramétriques

2018-2019

Objectif du test

Soient X et Y deux variables aléatoires quantitatives **continues**. On dispose de deux échantillons : $X_1, \ldots, X_n$ i.i.d e même loi que X et $Y_1, \ldots, Y_n$ i.i.d e même loi que Y. Le but de ce chapitre est de :

- Déterminer s'il existe une relation entre X et Y
- Caractériser la forme de la liaison (positive, négative; linéaire, monotone)
- Quantifier l'intensité de la relation
- Tester si la liaison est statistiquement significative

Remarque : on ne se place pas ici dans le cadre du modèle de régression linéaire où le but est d'évaluer l'influence d'une variable sur une autre.

Exemple sur des données réelles

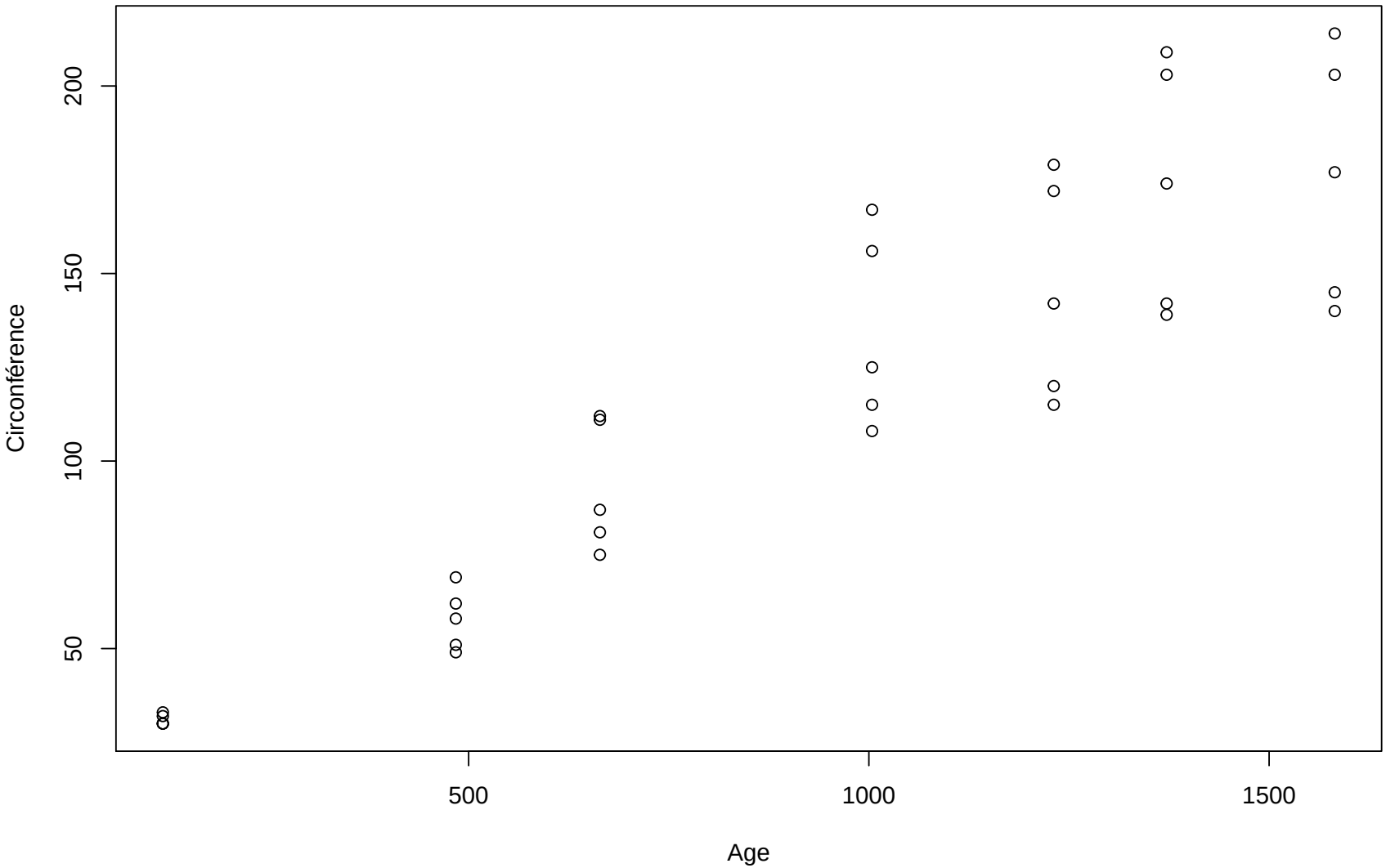
La base de données Orange de R contient des informations sur 35 orangers. On s'intéresse aux 2 variables suivantes :

- "age" représente l'âge des arbres, c'est une variable quantitative continue. Elle est mesurée en jours, l'arbre le plus jeune a 118 jours et le plus vieux 1582 jours.
- "circumference" représente la circonférence de l'arbre, c'est une variable quantitative continue, mesurée en mm. la plus petite circonférence est de 30 mm et la plus grosse de 214 mm.

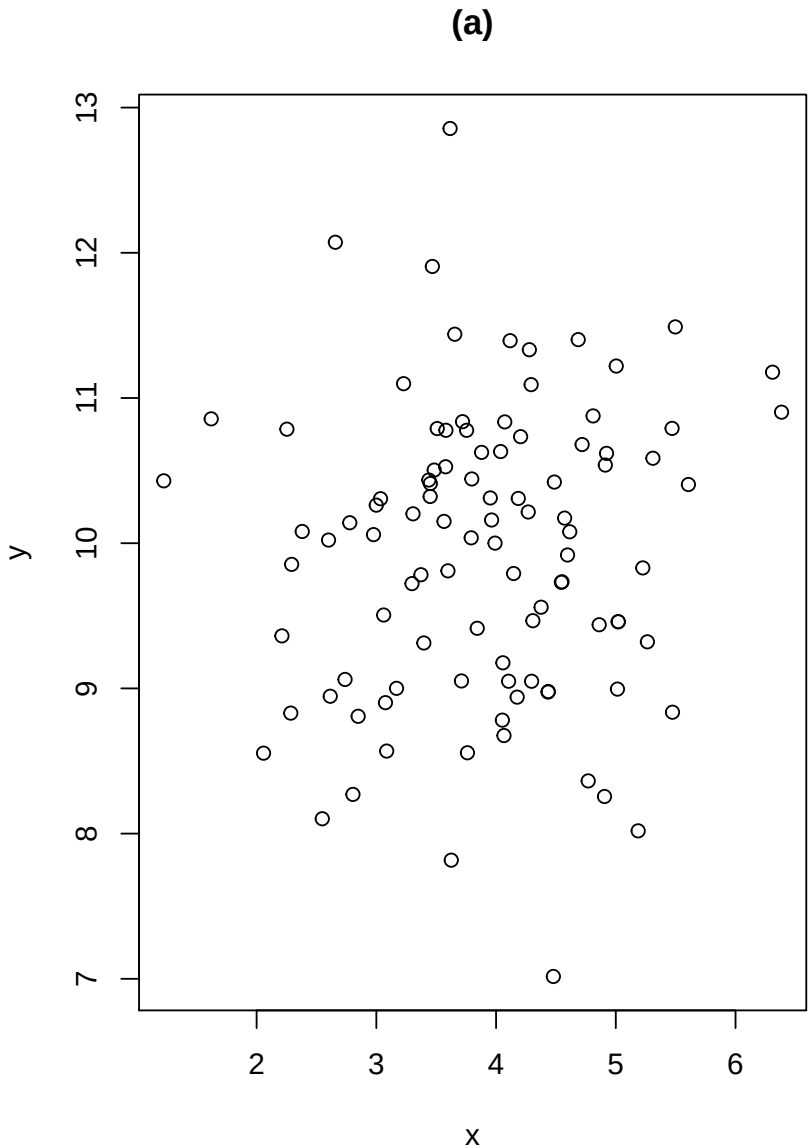
On s'intéresse au type de liason entre l'âge des arbres et la circonférence des arbres. Y-a-t-il une liaison entre ces deux variables ? Si oui, dans quel sens ? Peut-on quantifier l'intensité de la liaison ?

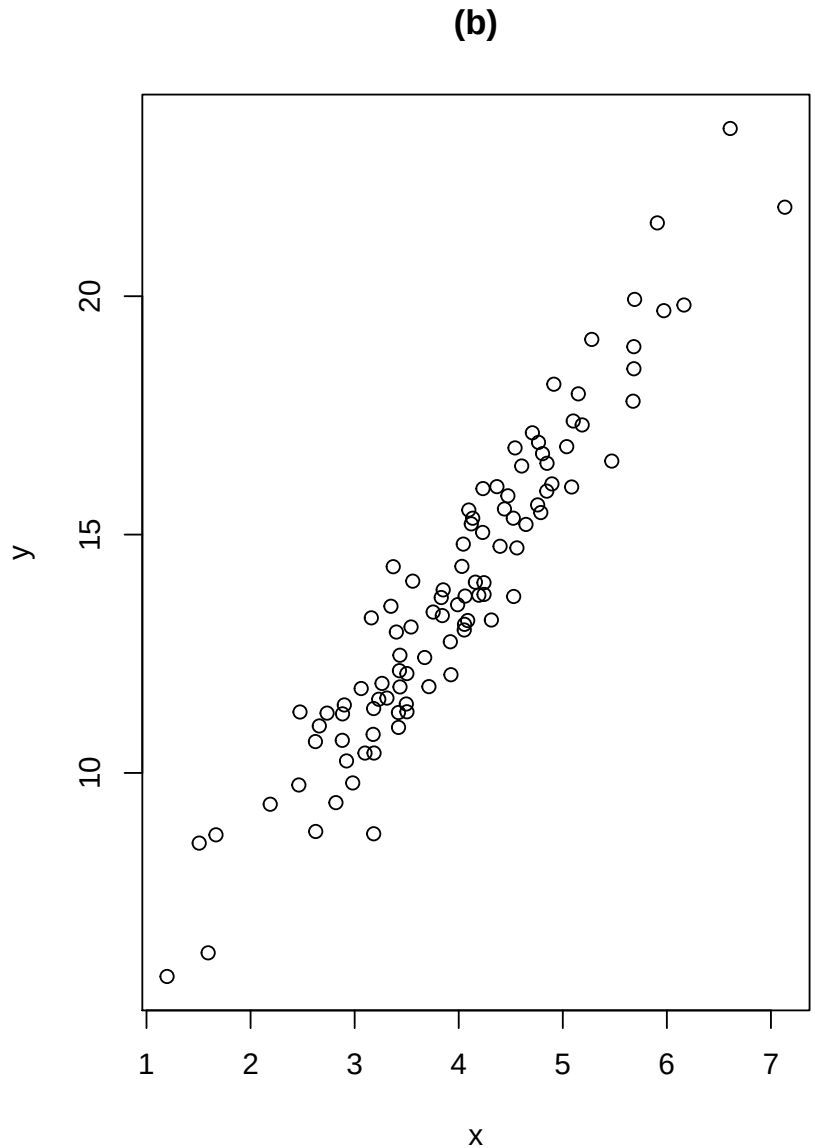
Exemple sur des données réelles (suite)

```
with(Orange,plot(age,circumference,xlab="Age",ylab=
"Circonférence"))
```

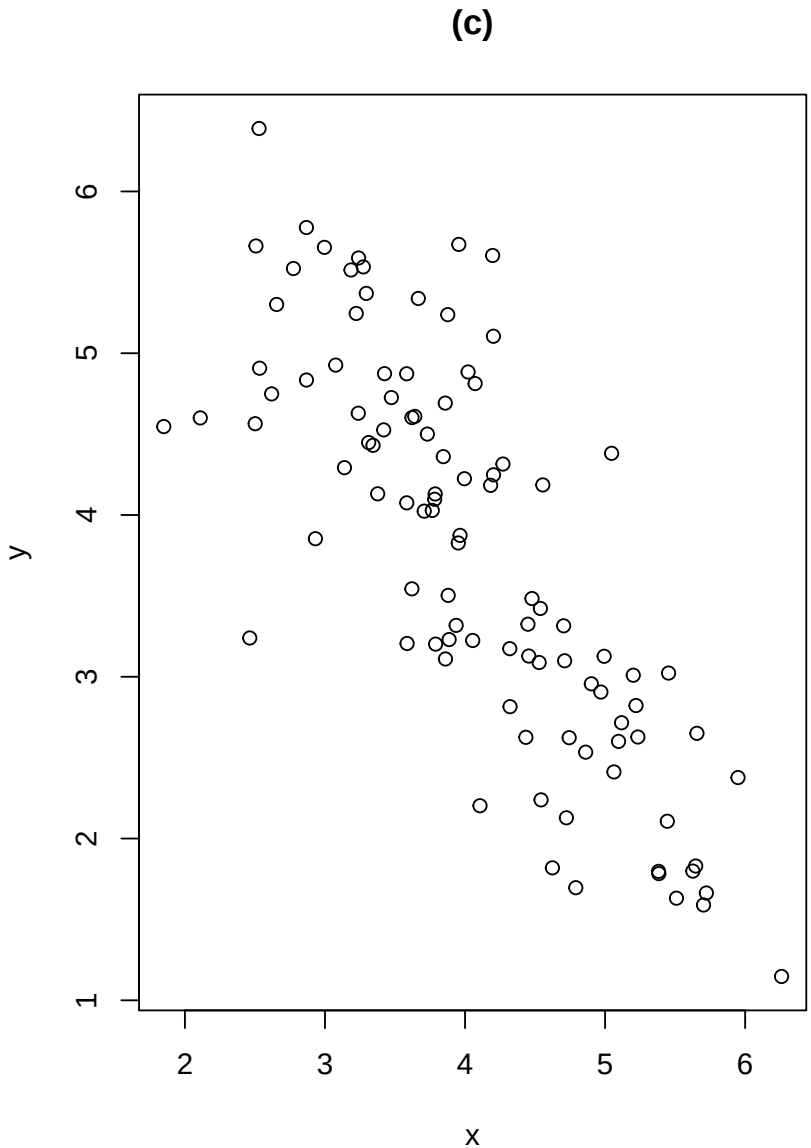


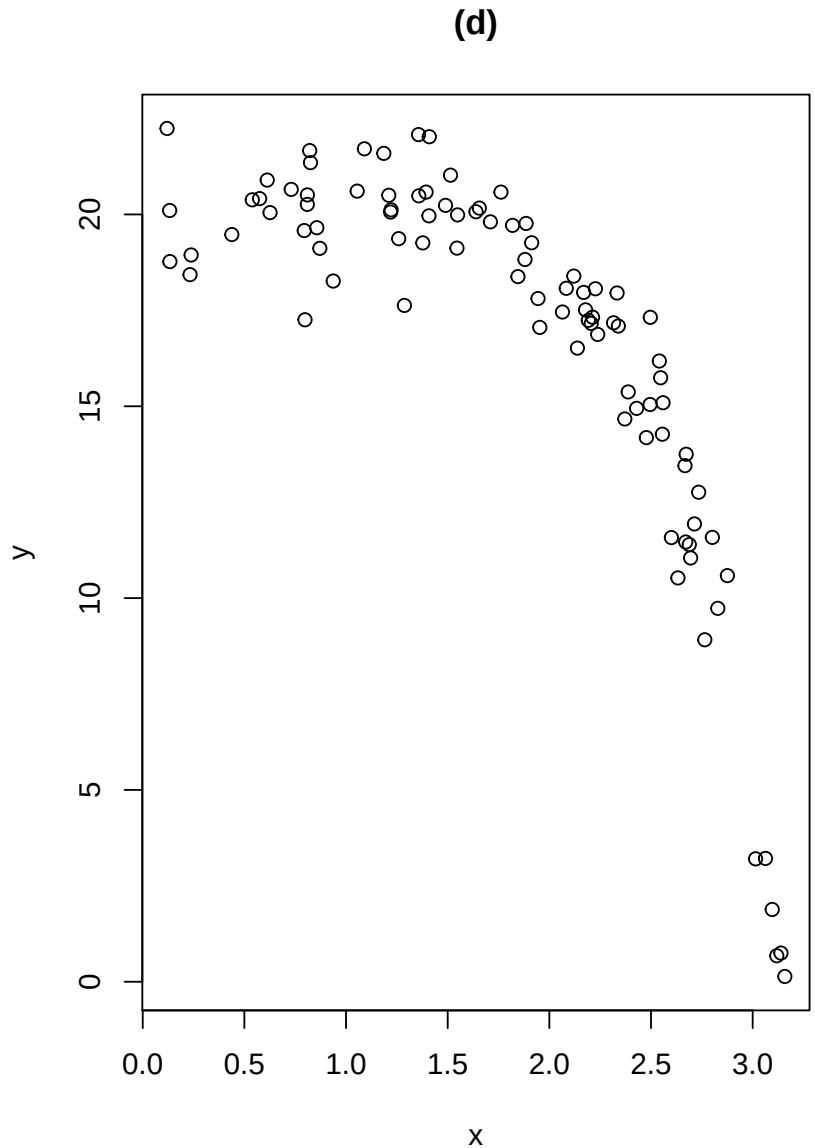
Exemples de liaisons entre deux variables



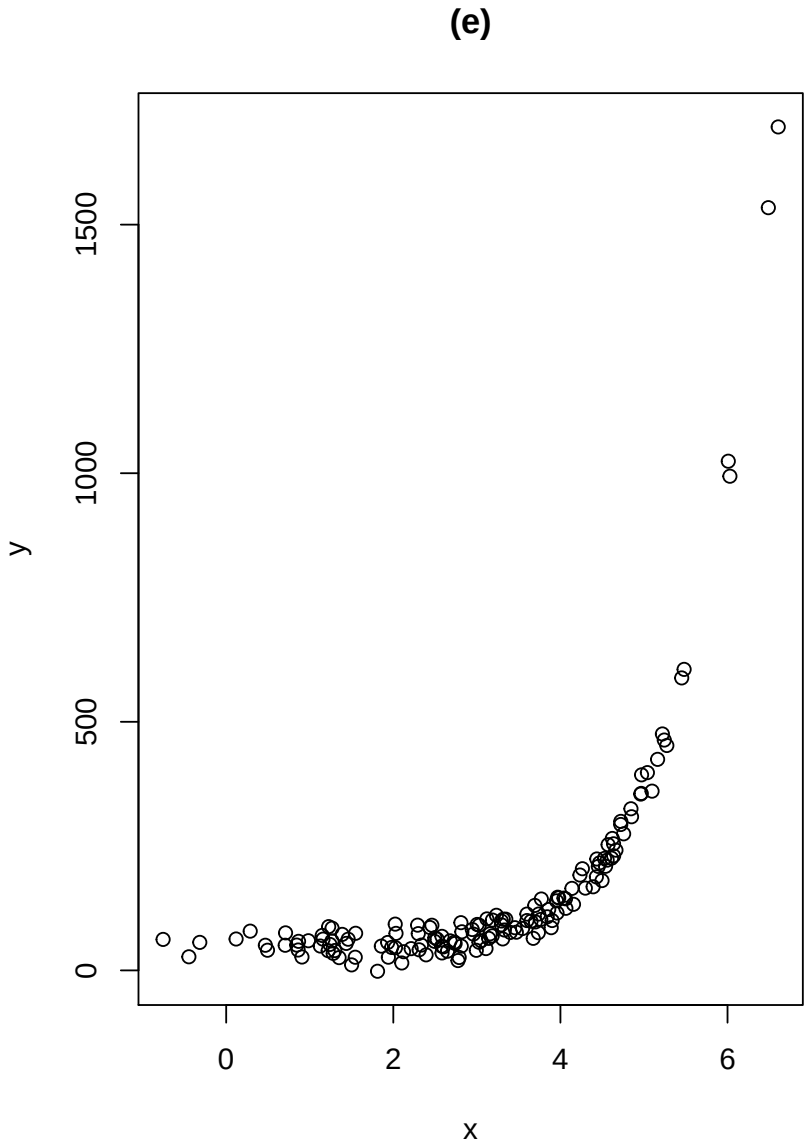


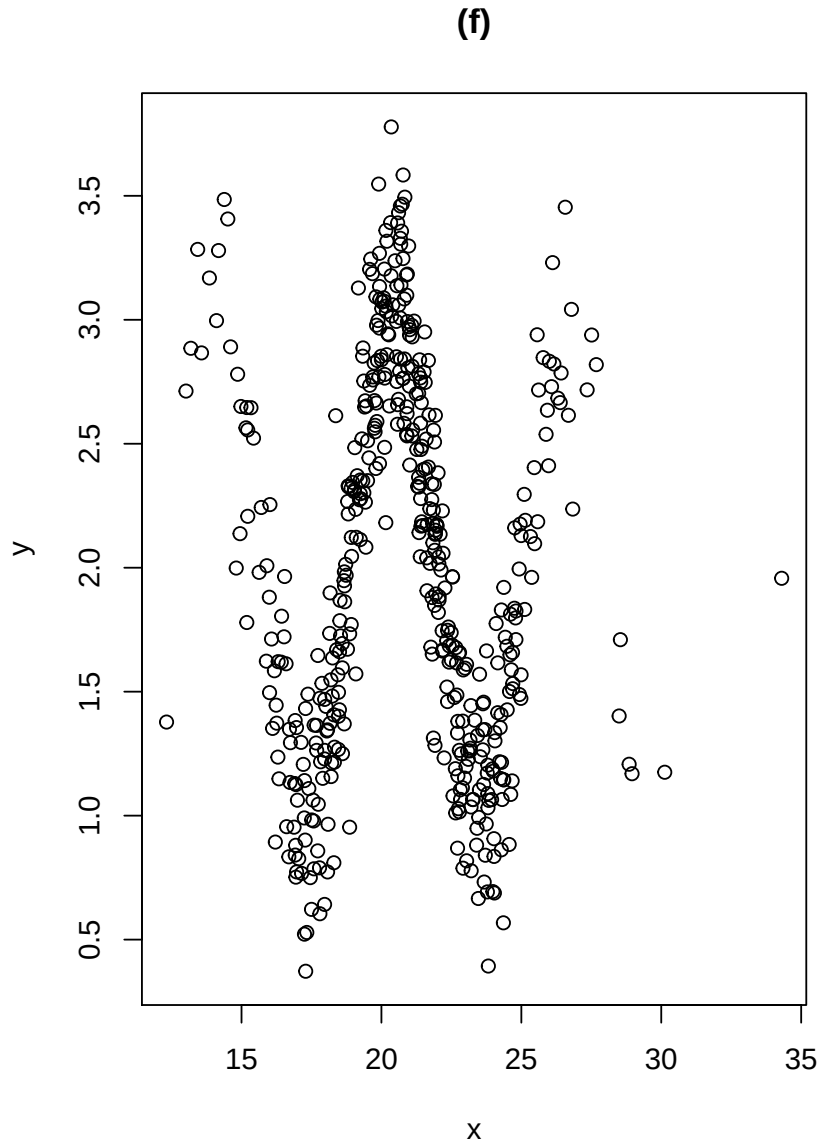
Exemples (suite)





Exemples (suite)





Exemples (suite)

- **(a)** Absence de liaison entre X et Y.
- **(b)** Liaison linéaire positive. Si X augmente alors Y aussi.
- **(c)** Liaison linéaire négative. Si X augmente alors Y diminue.
- **(d)** Liaison monotone négative mais non linéaire. Si X augmente alors Y diminue.
- **(e)** Liaison monotone positive mais non linéaire. Si X augmente alors Y aussi.
- **(f)** Liaison non linéaire et non monotone.

Mesures de la liaison entre X et Y

Il existe plusieurs mesures de liaison entre variables quantitatives continues. Nous utiliserons le coefficient de corrélation de **Pearson** et le coefficient de **Spearman**.

Le coefficient de corrélation de **Pearson** permet de mesurer le degré d'association pour des liaisons linéaires uniquement. Le coefficient de **Spearman** fonctionne également pour des liaisons monotones.

Coefficient de corrélation de Pearson

On rappelle la définition de la covariance entre X et Y :

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

et de la corrélation entre X et Y :

$$Cor(X, Y) = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}}$$

On a la propriété : $-1 \leq Cor(X, Y) \leq 1$.

Coefficient de corrélation de Pearson

La valeur du coefficient de corrélation de Pearson mesure le type de liaison linéaire entre X et Y.

- Si $Cor(X, Y) > 0$ il y a une liaison linéaire positive entre X et Y.
- Si $Cor(X, Y) < 0$ il y a une liaison linéaire négative entre X et Y.
- Si $Cor(X, Y) = 0$ il n'y a pas de liaison linéaire entre X et Y.

Remarque : $X \perp\!\!\!\perp Y \Rightarrow Cor(X, Y) = 0$ mais la réciproque est généralement fausse ! On a équivalence si le vecteur (X, Y) est un vecteur **gaussien**.

Estimation du coefficient de corrélation de Pearson

Le coefficient de corrélation **empirique** est un estimateur de $Cor(X, Y)$. Il est défini par :

$$\hat{r} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{j=1}^{n}(Y_j - \bar{Y})^2}}$$

Sous R on trouve :

- **(a)**

```
## [1] 0.07343106
```

Estimation du coefficient de corrélation de Pearson (suite)

- **(b)**

```
## [1] 0.9525128
```

- **(c)**

```
## [1] -0.7879123
```

- **(d)**

```
## [1] -0.7631126
```

Estimation du coefficient de corrélation de Pearson (suite)

- **(e)**

```
## [1] 0.6525232
```

- **(f)**

```
## [1] -0.05069953
```

Lien avec le modèle linéaire

- Le coefficient de corrélation est proportionnel à la pente de la droite des moindres carrés dans le modèle linéaire !
- Le coefficient $\hat{r}^2$ s'interprète comme la proportion de variance de Y qui est linéairement expliquée par X. Il s'appelle le coeficient de **détermination**.

Par exemple, pour le cas **(b)**, on peut dire au vue du coefficient de corrélation que la liaison entre X et Y est forte. Par ailleurs, R nous donne le coefficient de détermination suivant :

```
## [1] 0.9072807
```

Limites du coefficient de corrélation de Pearson

- Dans les exemples **(d)** et **(e)** la liaison est monotone mais non linéaire : le coefficient de corrélation donne des indications sur l'existence de liaison entre X et Y mais traduit mal son **intensité**.
- Dans l'exemple **(f)**, la liaison n'est n'y monotone ni linéaire : le coefficient de corrélation de Pearson n'est pas adapté !

Le coefficient de Spearman

Le coefficient de Spearman est un coefficient de corrélation basé sur les **rangs** des observations. On note, pour $i = 1, \ldots, n$, R_i le rang de X_i au sein de l'échantillon global $X_1, \ldots, X_n$ et S_i le rang de Y_i au sein de l'échantillon global $Y_1, \ldots, Y_n$

Par exemple, pour $n = 7$, si les réalisations de $X_1, \ldots, X_7$ sont :

```
## [1] 6 1 8 9 3 7 2
```

alors, les réalisations de $R_1, \ldots, R_7$ sont :

```
## [1] 4 1 6 7 3 5 2
```

Le coefficient de Spearman (suite)

Le coefficient de Spearman est un coefficient de corrélation de Pearson calculé sur les rangs des deux échantillons. On le note ρ et $\hat{\rho}$ sa version empirique. On a :

$$\rho = Cor(R, S)$$

et

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{j=1}^{n}(S_j - \bar{S})^2}}$$

Le coefficient de Spearman (suite)

Ce coefficient s'interpréte de manière similaire au coefficient de Pearson : $-1 \leq \rho \leq 1$.

- Si $\rho > 0$ il y a une liaison **monotone** positive entre X et Y.
- Si $\rho < 0$ il y a une liaison **monotone** négative entre X et Y.
- Si $\rho = 0$ il n'y a pas de liaison **monotone** entre X et Y.

Par ailleurs, $X \perp\!\!\!\perp Y \Rightarrow \rho = 0$.

Par contre, ρ^2 n'a pas d'interprétation statistique !

Le coefficient de Spearman (suite)

Ce coefficient a donc l'avantage de pouvoir caractériser des liaisons non linéaires mais monotones.

Sous R on trouve :

- **(d)**

```
## [1] -0.8596551
```

- **(e)**

```
## [1] 0.8659638
```

Le coefficient de Spearman (suite)

Le coefficient de Spearman est par ailleurs **robuste** aux points atypiques.

Par contre, il ne permet pas de caractériser une liaison non linéaire et non monotone !

Sous R on trouve :

- **(f)**

```
## [1] -0.06400384
```

Test de corrélation de Pearson

Le test de corrélation de Pearson teste les hypothèses :

$$(H_0) Cor(X, Y) = 0 \quad (H_1) Cor(X, Y) \neq 0$$

La statistique de test utilisée est :

$$T_n = \frac{\hat{r}}{\sqrt{(1 - \hat{r}^2)/(n-2)}}$$

Statistique de test dans le cas gausien et région de rejet

- Sous (H_0), si le vecteur (X, Y) suit une loi normale bivariée alors T_n suit une loi de Student à $n-2$ degrés.
- La région de rejet s'écrit : $R_\alpha = \{|T_n| > c_\alpha\}$ où c_α est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-2$ degrés.

L'hypothèse sur la loi de (X, Y) revient à tester l'indépendance entre X et Y. C'est une hypothèse impossible à vérifier en pratique ! On utilisera plutôt la version asymptotique de ce test.

Statistique de test dans le cas asymptotique et région de rejet

- Si n est suffisament "grand", on peut approximer la statistique de test par une loi gaussienne. Sous (H_0), T_n suit approximativement une loi normale centrée réduite.
- La région de rejet s'écrit : $R_\alpha = \{|T_n| > c_\alpha\}$ où c_α est le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

Test de Spearman

Le test de Spearman est un test **exact** qui teste les hypothèses :

$$(H_0) Cor(R, S) = 0 \quad (H_1) Cor(R, S) \neq 0$$

La statistique de test utilisée est :

$$T_n = \frac{\hat{\rho}}{\sqrt{(1 - \hat{\rho}^2)/(n-2)}}$$

Elle suit une loi de Student à $n-2$ degrés de liberté sous (H_0). Ce résultat est valide sans faire d'hypothèse sur la loi de (X, Y) ! Pour n grand, la statistique de test s'approxime par une loi normale centrée réduite sous (H_0).

Retour aux exemples **(a)** : test de Pearson

```
##
##  Pearson's product-moment correlation
##
## data:  x1 and y1
## t = 0.7289, df = 98, p-value = 0.4678
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  -0.1247869  0.2660124
## sample estimates:
##       cor
## 0.07343106
```

Retour aux exemples **(a)** : test de Spearman

```
##
##  Spearman's rank correlation rho
##
## data:  x1 and y1
## S = 151710, p-value = 0.3745
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.08963696
```

Retour aux exemples **(e)** : test de Pearson

```
##
##  Pearson's product-moment correlation
##
## data:  x5 and y5
## t = 10.476, df = 148, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.5497516 0.7358352
## sample estimates:
##       cor
## 0.6525232
```

Retour aux exemples **(e)** : test de Spearman

```
##
##  Spearman's rank correlation rho
##
## data:  x5 and y5
## S = 75392, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8659638
```

Retour aux exemples **(f)** : test de Pearson

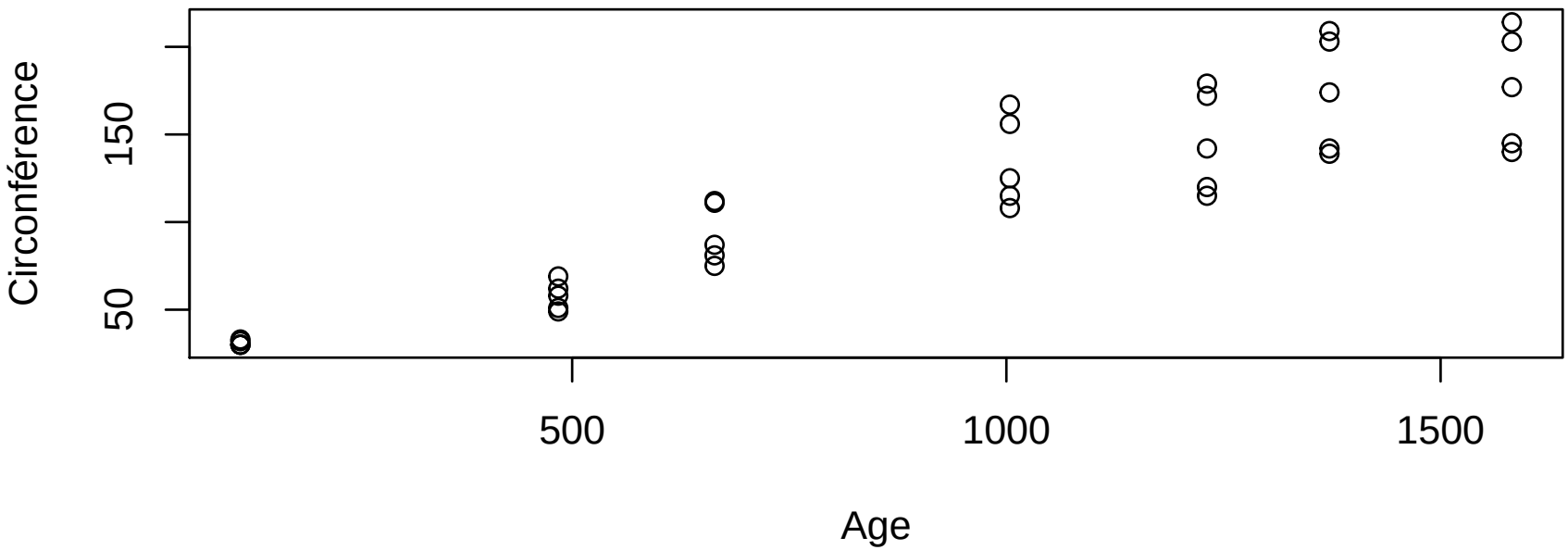
```
##
##  Pearson's product-moment correlation
##
## data:  x6 and y6
## t = -1.1329, df = 498, p-value = 0.2578
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  -0.13777758  0.03715625
## sample estimates:
##         cor
## -0.05069953
```

Retour aux exemples **(f)** : test de Spearman

```
##
##  Spearman's rank correlation rho
##
## data:  x6 and y6
## S = 22167000, p-value = 0.1529
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##         rho
## -0.06400384
```

Etude de cas : retour sur les données d'oranger

```
with(Orange,plot(age,circumference,xlab="Age",ylab=
"Circonférence"))
```



Les données d'oranger (suite)

```
with(Orange,cor.test(age,circumference))
```

```
##
##  Pearson's product-moment correlation
##
## data:  age and circumference
## t = 12.9, df = 33, p-value = 1.931e-14
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.8342364 0.9557955
## sample estimates:
##       cor
## 0.9135189
```

Les données d'oranger (suite)

```
with(Orange,cor.test(age,circumference,method="spearman"))
```

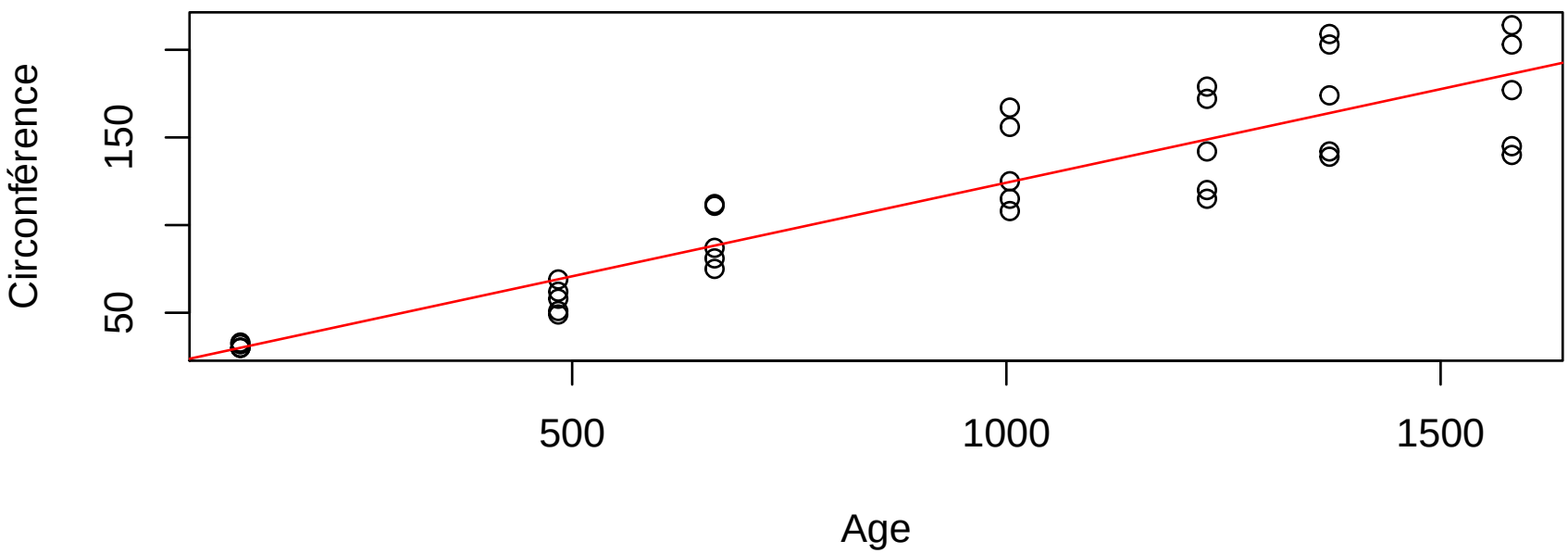
```
## Warning in cor.test.default(age, circumference, method =
## Cannot compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  age and circumference
## S = 668.09, p-value = 6.712e-14
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9064294
```

Les données d'oranger (suite)

Le coefficient de corrélation est relié à la pente de la droite des moindres carrés.

```
modelOrange<-lm(circumference~age,data=Orange)
with(Orange,plot(age,circumference,xlab="Age",ylab=
"Circonférence"))
abline(modelOrange,col="red")
```



Etude de cas sur les graines de soja

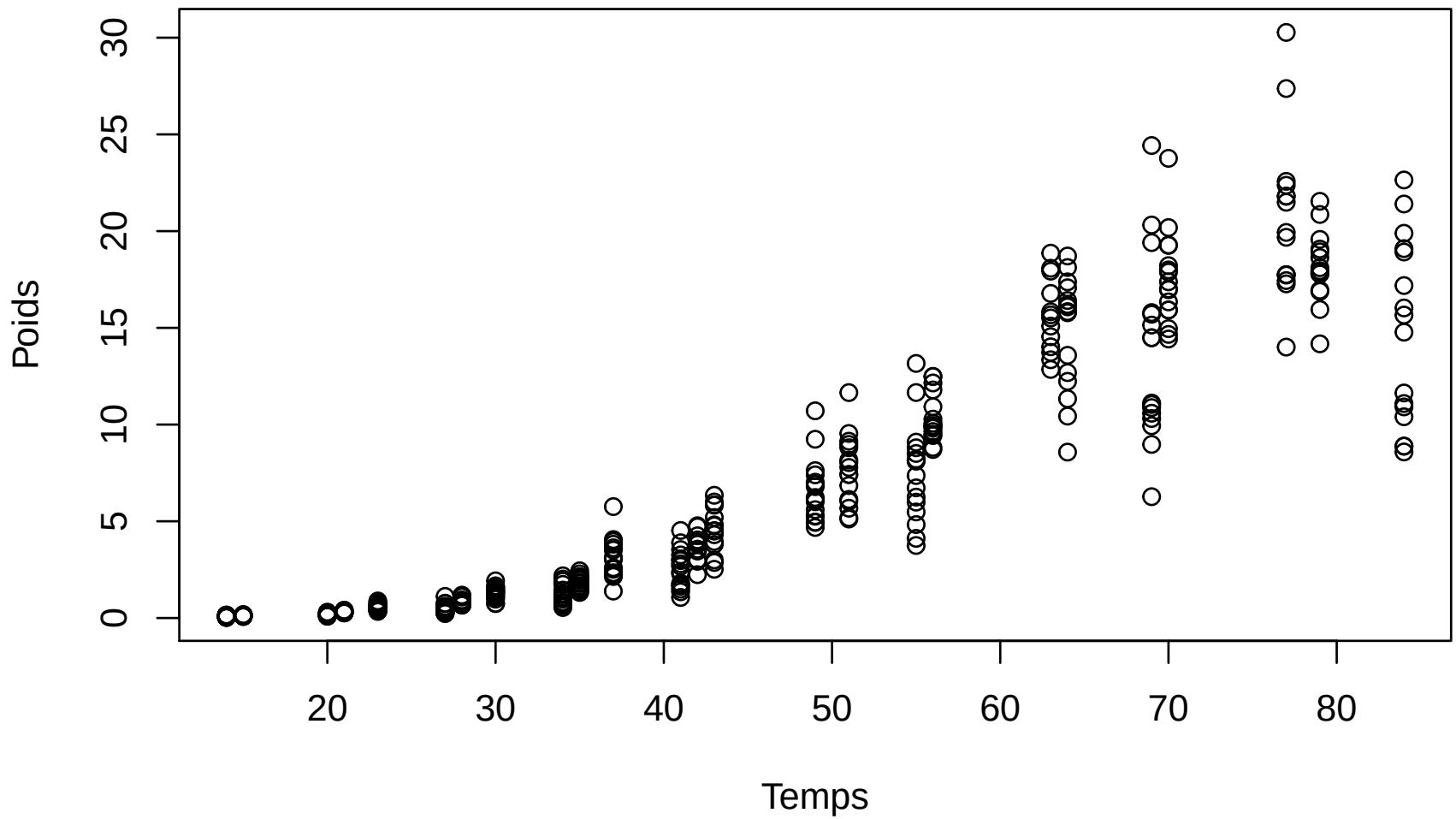
La base de données Soybean de R contient des informations sur 412 graines de soja qui ont été plantées. On s'intéresse aux 2 variables suivantes :

- "Time" représente le jour où la plante de soja est analysée.
- "weight" représente le poids en grammes de la feuille de soja.

On s'intéresse à la liaison entre ces deux variables. Est-ce que le poids des feuilles de soja évolue à mesure que le temps passe ? Si oui dans quel sens ?

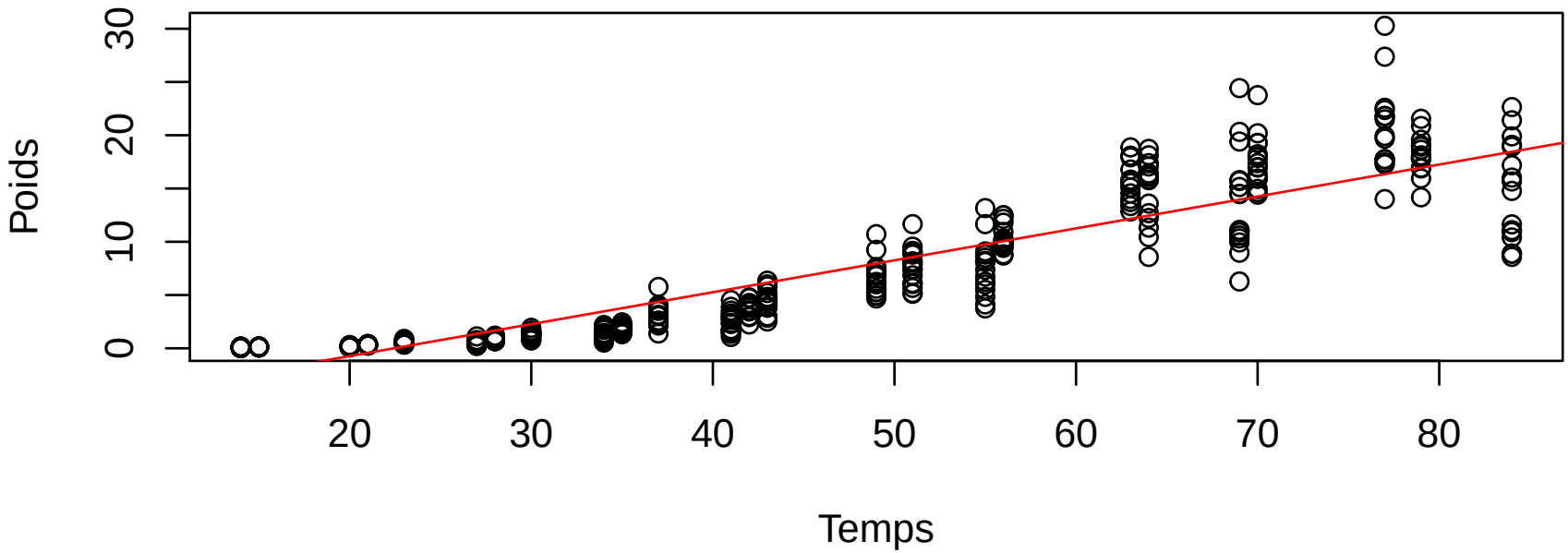
Les graines de soja (suite)

```
library(nlme)
with(Soybean,plot(Time,weight,xlab="Temps",ylab="Poids"))
```



Les graines de soja (suite)

```
modelSoy<-lm(weight~Time,data=Soybean)
with(Soybean,plot(Time,weight,xlab="Temps",ylab="Poids"))
abline(modelSoy,col="red")
```



La droite de régression ne caractérise pas parfaitement la relation entre le temps et le poids !

Les graines de soja (suite)

```
with(Soybean,cor.test(Time,weight))
```

```
##
##  Pearson's product-moment correlation
##
## data:  Time and weight
## t = 44.707, df = 410, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  0.8928916 0.9260385
## sample estimates:
##       cor
## 0.9109236
```

Les graines de soja (suite)

```
with(Soybean,cor.test(Time,weight,method="spearman"))
```

```
## Warning in cor.test.default(Time, weight, method = "spea
## compute exact p-value with ties


##
##  Spearman's rank correlation rho
##
## data:  Time and weight
## S = 314500, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.9730172
```