Statistique descriptive bidimensionnelle

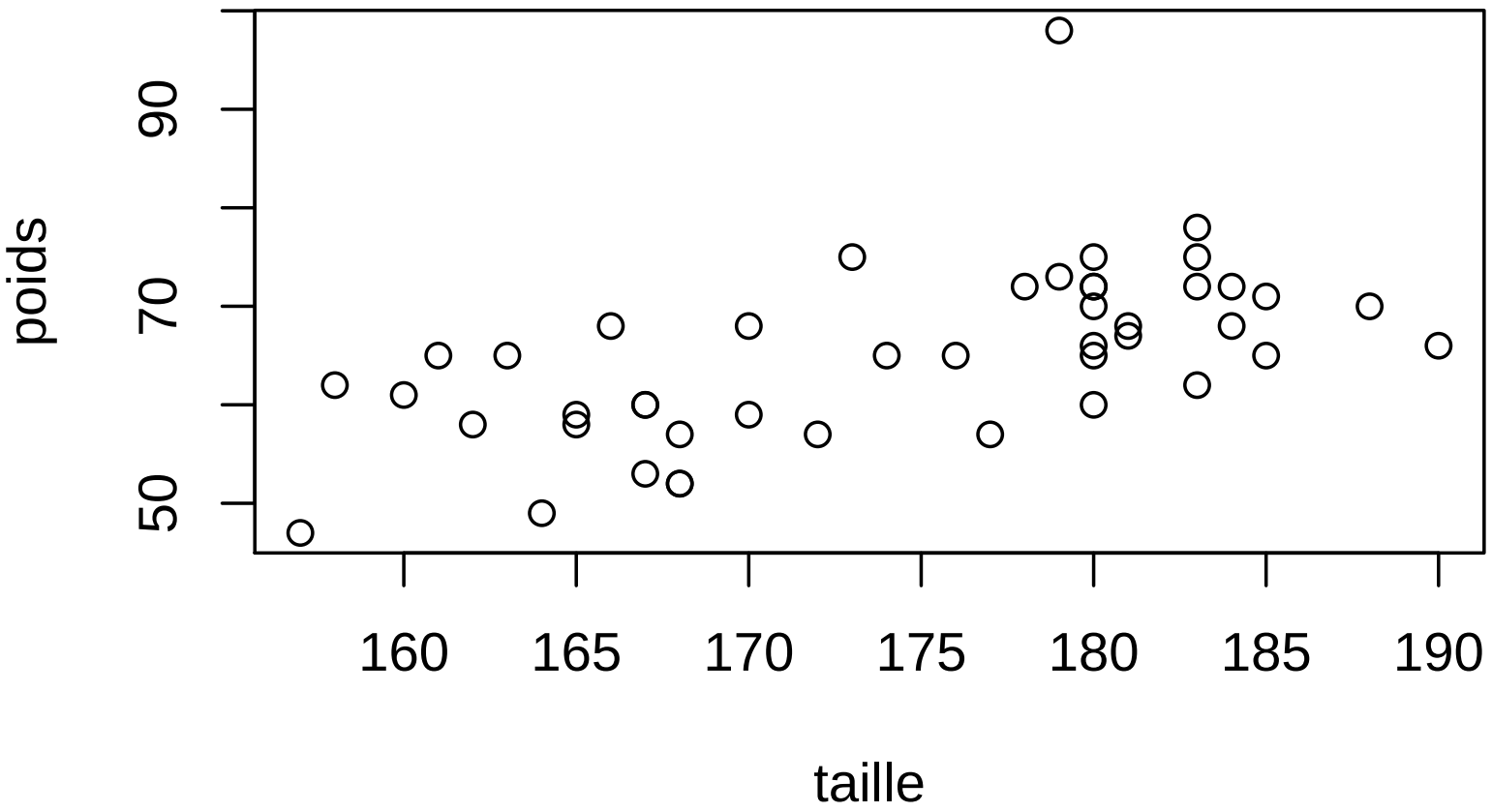
Antonio Falcó

1. Diagramme de dispersion
2. Covariance et corrélation
3. Régresion linéaire
4. Exemple pratique

Comme dans la lesson précédent, nous allons nous concentrer sur les variables quantitatives avec un grand nombre de modalités.

Pour visualiser l'association entre deux telles variables, le moyen le plus simple est de construire un diagramme de dispersion ou *scatter plot*. Un diagramme de dispersion représente les observations de deux variables en reportant l'une sur l'axe horizontal et l'autre sur l'axe vertical (nuage de points).

Voici par exemple le diagramme de dispersion des poids et tailles des étudiant(e)s de première année:



Le graphique semble indiquer une association entre les variables poids et taille: une plus grande taille semble correspondre en moyenne à un plus grand poids.

Une façon de quantifier cette association est le coefficient de covariance. Pour deux variables X et Y **mesurées sur les mêmes unités d'observation**, le coefficient de covariance (ou simplement covariance), noté $v(X, Y)$, est défini par:

$$v(X, Y) = m((X - m(X))(Y - m(Y))) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m(X))(y_i - m(Y)).$$

Propriétés de la covariance

Soient X, Y et Z des variables et soient a, b, c et d des constantes.

1. Si $v(X, Y) > 0$, cela suggére que les grandes valeurs de X sont généralement associées aux grandes valeurs de Y et les petites valeurs de X aux petites valeurs de Y (la présence d'outliers peut invalider ces interprètations).
2. Si $v(X, Y) < 0$, cela suggére que les grandes valeurs de X sont généralement associées aux petites valeurs de Y et les petites valeurs de X aux grandes valeurs de Y (la présence d'outliers peut invalider ces interprètations).

3. $v(X, X) = s^2(X)$
4. Symétrie: $v(X, Y) = v(Y, X)$
5. $v(X, c) = 0$
6. $v(aX + bY, Z) = a\, v(X, Z) + b\, v(Y, Z)$
7. $v(aX + b, cY + d) = ac\, v(X, Y)$
8. $s^2(X + Y) = s^2(X) + s^2(Y) + 2v(X, Y)$
9. $v(X, Y) = m(XY) - m(X)m(Y)$

La propriété 9. est pratique pour faire le calcul à la main car elle évite de calculer tous les écarts $(x_i - m(X))$ et $(y_i - m(Y))$.

L'inconvénient de la covariance comme mesure de l'association entre deux variables est qu'elle **dépend des unités de mesures**.

Par exemple, la covariance entre les tailles et les poids des étudiant(e)s vaut $v(T, P) = 41.82$ cm kg. Si on d'ecidait de mesurer la taille en métres (Tm) et le poids en grammes (Pg), on obtiendrait $v(Tm, Pg) = 418.2$ m g. Or, il est clair que l'association entre la taille et le poids des ètudiants ne d'epend pas des unités dans lesquelles elles sont mesurées!

Il est donc difficile d'interpréter la covariance entre deux variables.

Par contre, si on prend les tailles et les poids standarisées:

$$T_s = \frac{T - m(T)}{s(T)} \text{ et } P_s = \frac{P - m(P)}{s(P)}.$$

On obtient,

$$\begin{aligned} v(T_s, P_s) &= \frac{1}{s(T)s(P)} v((T - m(T)), (P - m(P))) \\ &= \frac{v(T, P)}{s(T)s(P)} \text{ (après 7.)} \end{aligned}$$

La covariance de les variables standarisées est une mesure sans unité.

Pour remédier à cet inconvénient, on définit le coefficient de corrélation (ou simplement corrélation), noté $r(X, Y)$, entre les variables X et Y comme

$$r(X, Y) = \frac{v(X, Y)}{s(X)s(Y)} = v\left(\frac{X - m(X)}{s(X)}, \frac{Y - m(Y)}{s(Y)}\right)$$

Pour les poids et tailles, on obtient

$$r(T, P) = r(Tm, Pg) = 0.64.$$

La corrélation est une mesure sans unité. Elle est donc interprétable même dans des cas où les unités des variables ne nous sont pas familières.

Propriétés de la corrélation

Soient X, Y et Z des variables et soient a, b, c et d des constantes.

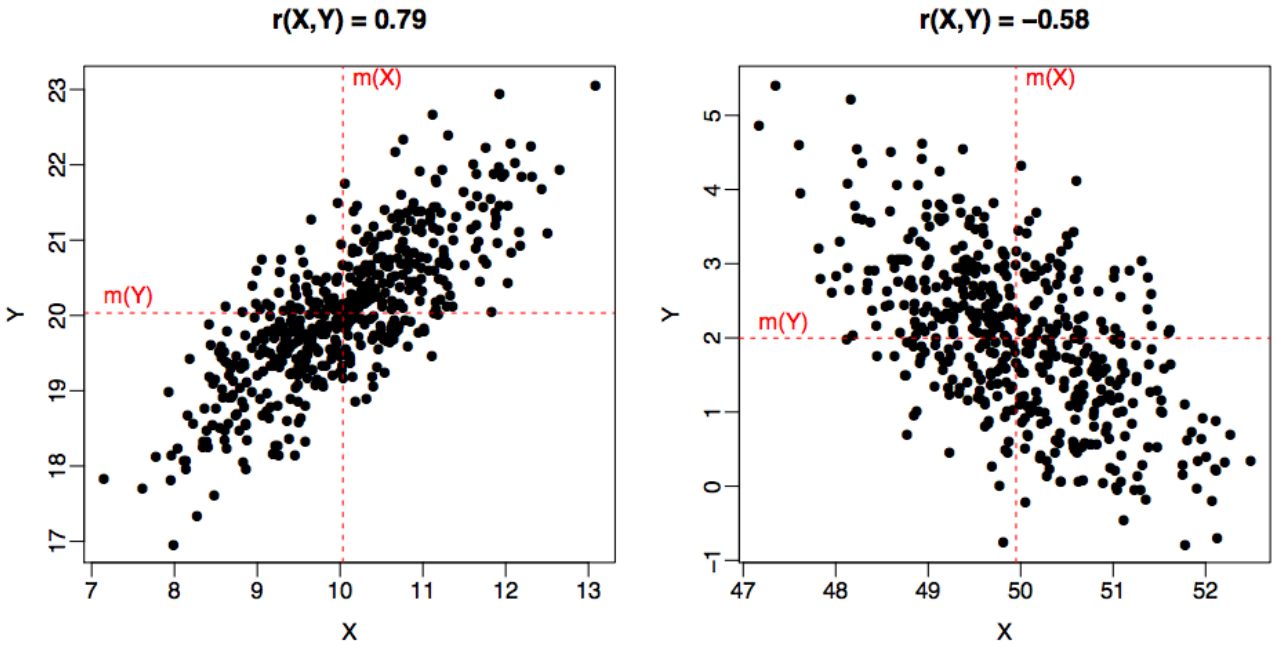
1. Si $r(X, Y) > 0$, cela suggére que les grandes valeurs de X sont généralement associées aux grandes valeurs de Y et les petites valeurs de X aux petites valeurs de Y (la présence d'outliers peut invalider ces interprètations).
2. Si $r(X, Y) < 0$, cela suggére que les grandes valeurs de X sont généralement associées aux petites valeurs de Y et les petites valeurs de X aux grandes valeurs de Y (la présence d'outliers peut invalider ces interprètations).

3. $r(X, X) = 1$
4. Symétrie: $r(X, Y) = r(Y, X)$
5. $r(X, c) = \frac{v(X,c)}{s(X)\cdot 0} = \frac{0}{0}$!!! est **indéfini**.
6. $r(aX + b, cY + d) = \text{signe}(ac)\, r(X, Y)$
7. $r(aX + b, X) = \text{signe}(a) r(X, X) = \pm 1.$
8. $-1 \leq r(X, Y) \leq 1.$

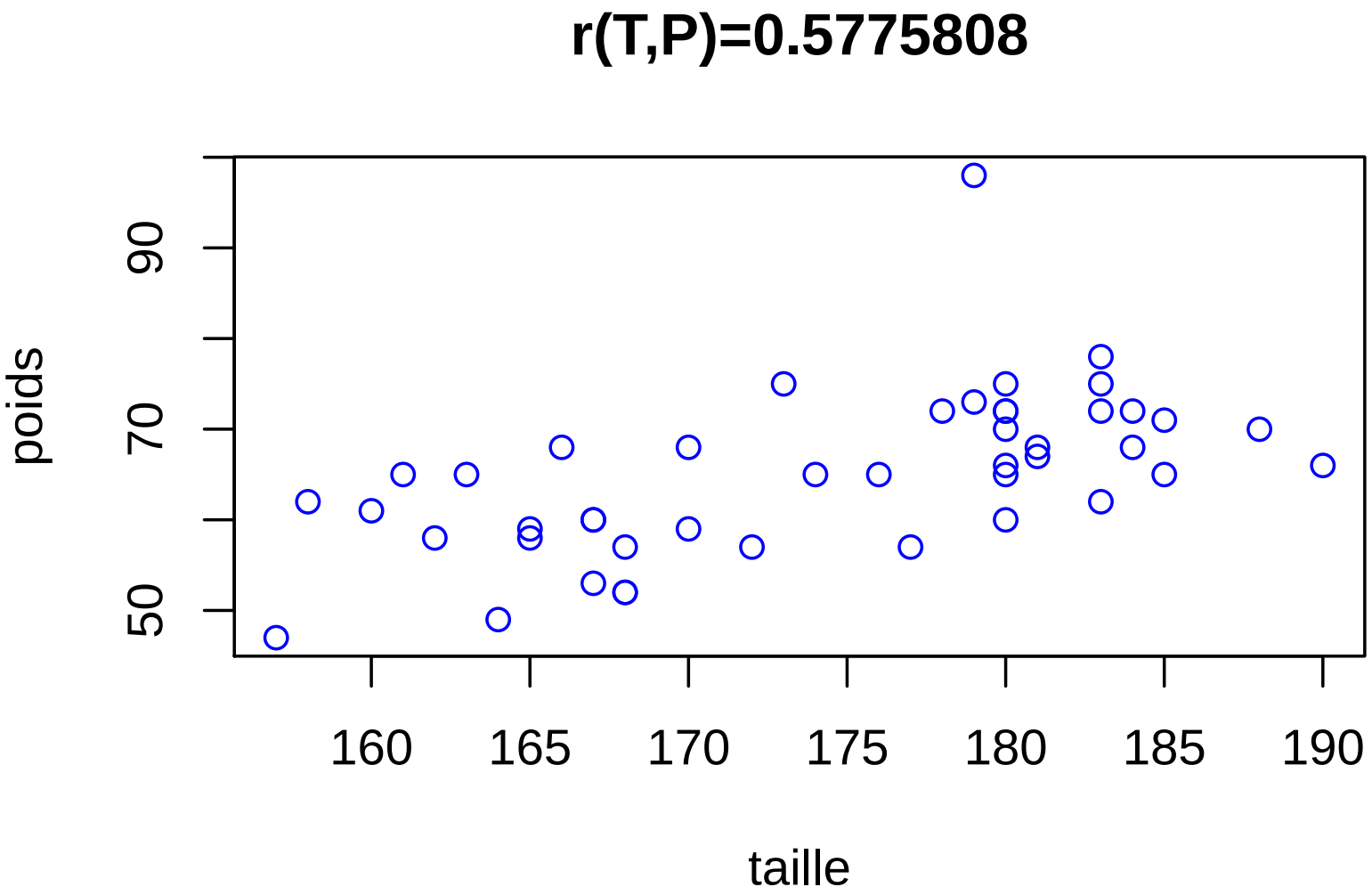
La corrélation entre deux variables est donc toujours comprise entre -1 et 1, et ces bornes maximale et minimale sont atteintes lorsqu'il a y une relation linéaire parfaite entre les variables.

La corrélation est une mesure de l'association linéaire entre deux variables.

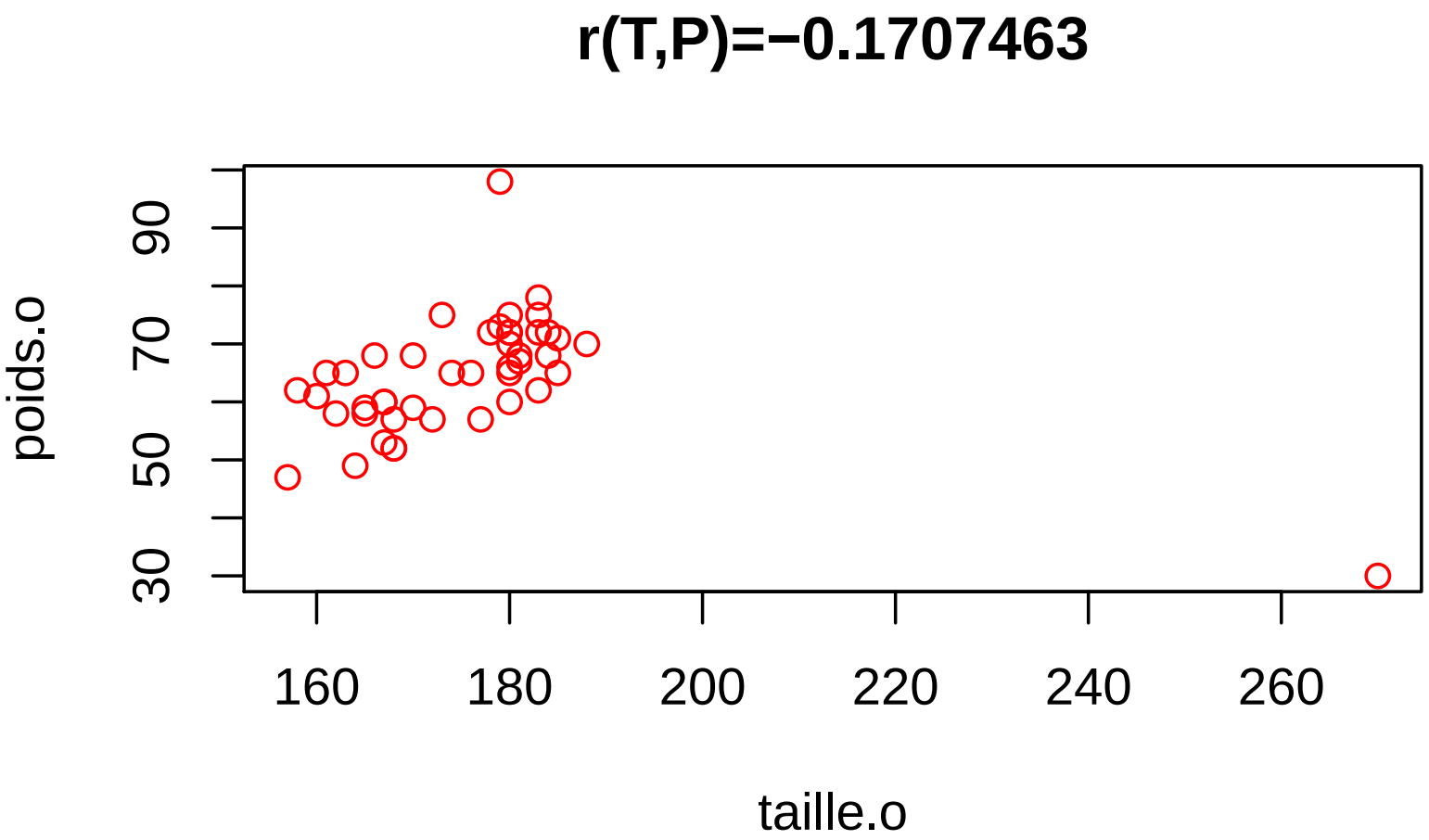
Une autre formulation des propriétés 1. et 2. est la suivante:



Si une valeur de X supérieure à la moyenne de X est généralement associée à une valeur de Y supérieure à la moyenne de Y, et de même pour les valeurs inférieures à la moyenne, $r(X, Y)$ aura tendance à être positif. Une association renversée conduira $r(X, Y)$ à être nêgatif.

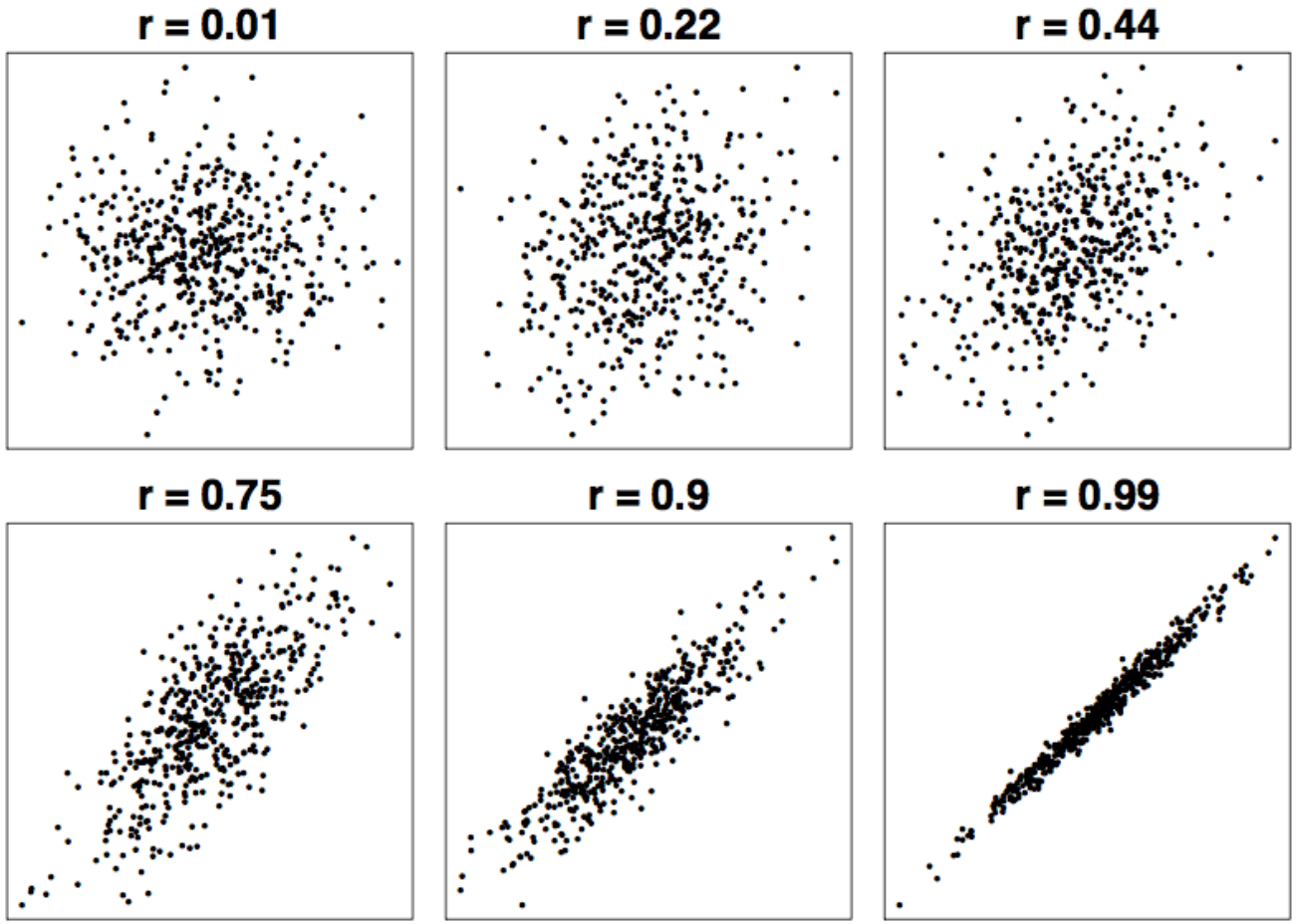

r(T,P)=0.5775808
poids
50
70
90
160
165
170
175
180
185
190
taille

Effet d'un outlier sur la corrélation Cet exemple montre que la présence d'un seul outlier peut complétement changer la valeur de la corrélation et invalider l'interprétation usuelle:

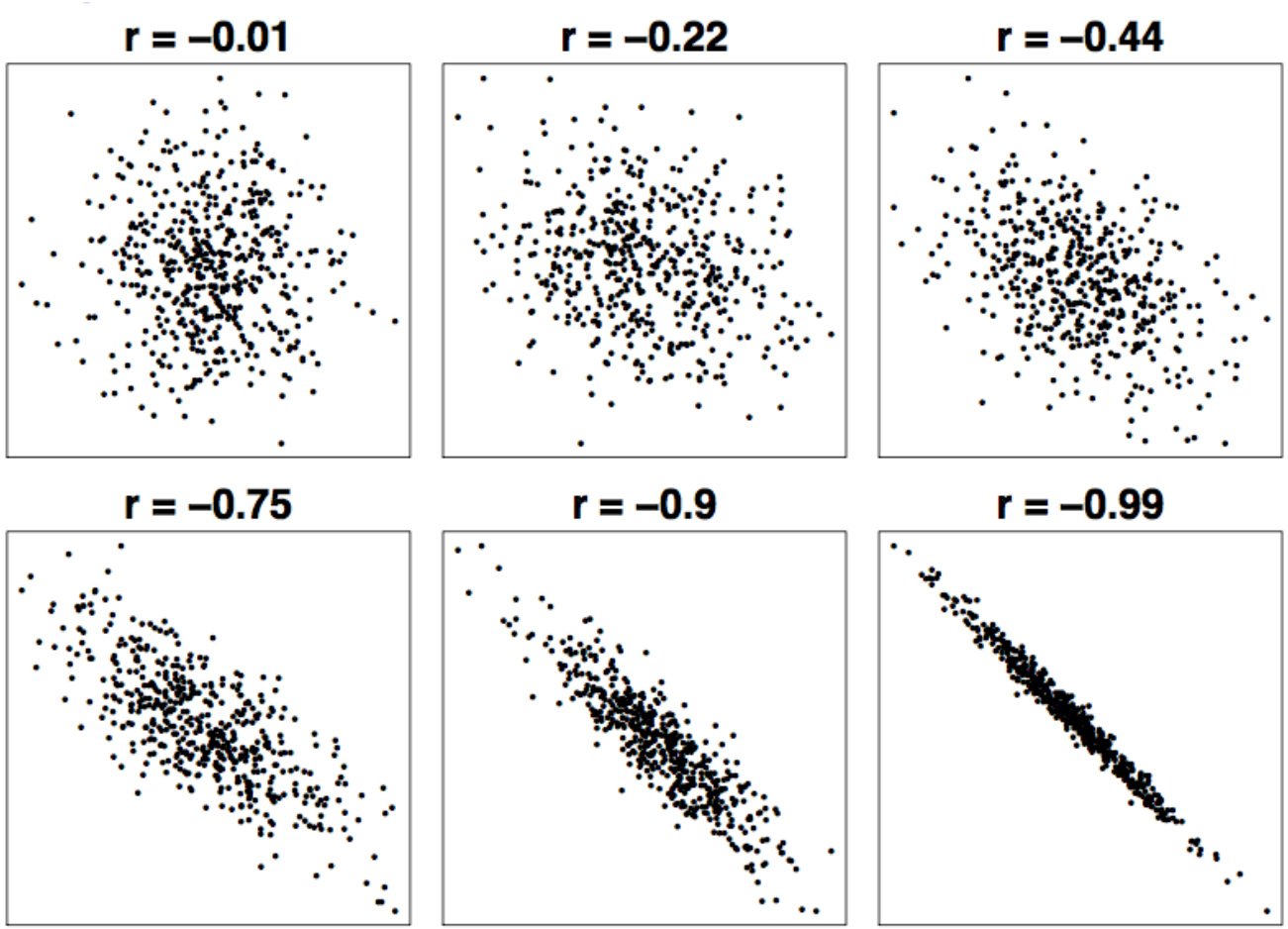


D'où l'importance de regarder les scatter plots avant d'interpréter une corrélation!

Voici quelques exemples de diagrammes de dispersion correspondant à différentes valeurs positives de la corrélation:

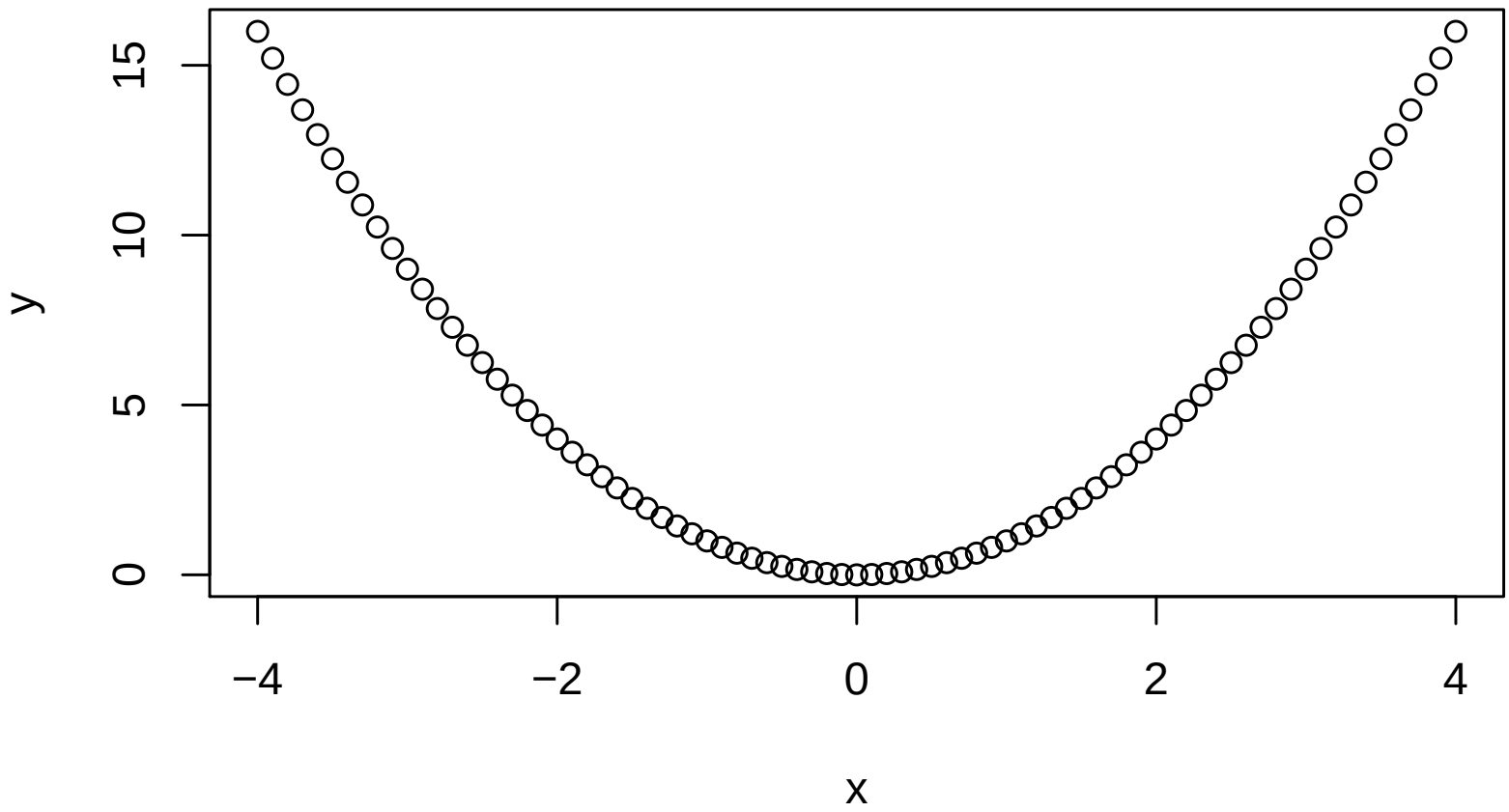


et voici quelques exemples de diagrammes de dispersion correspondant à différentes valeurs negatives de la corrélation:



Lorsqu'on interprète une corrélation, il convient d'être attentif aux points suivants:

- Une corrélation nulle ne signifie pas qu'il n'y a pas de relation entre deux variables, elle signifie seulement qu'il n'y a pas d'association linéaire. Par exemple dans le cas ci-dessous il y a une association quadratique exacte entre les deux variables, mais la corrélation est nulle.



$$r(X, Y) = 1.6336629 \times 10^{-16} \approx 0.$$

Pour d'ecrire de façon plus détaillée la relation entre deux variables X et Y, on cherche un modèle mathématique de cette relation, caractérisé par une fonction $y = f(x)$. L'avantage est qu'un modèle plus simple sera plus facile à interpréter. Un modèle très courant est celui de la régression linéaire, où la fonction $y = f(x)$ est une droite. On appelle cette droite la droite de régression.

- Quelle droite choisir?

Celle qui 'colle' le mieux aux données, selon un certain critère.

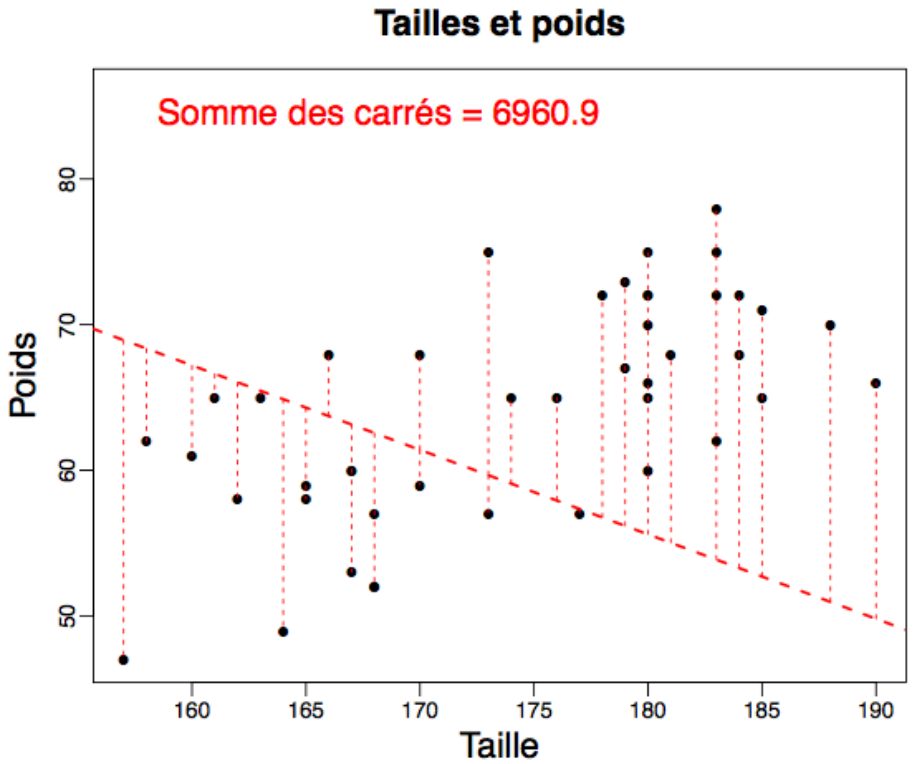
- Critère généralement utilisé: critère des moindres carrés.

Choisir la droite qui minimise la somme des carrés des distances entre la droite et les observations.

On cherche $\widehat{\beta}_0$ et $\widehat{\beta}_1$ telles que

$$\sum_{i=1}^{n}(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1\, x_i))^2 \leq \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1\, x_i))^2$$

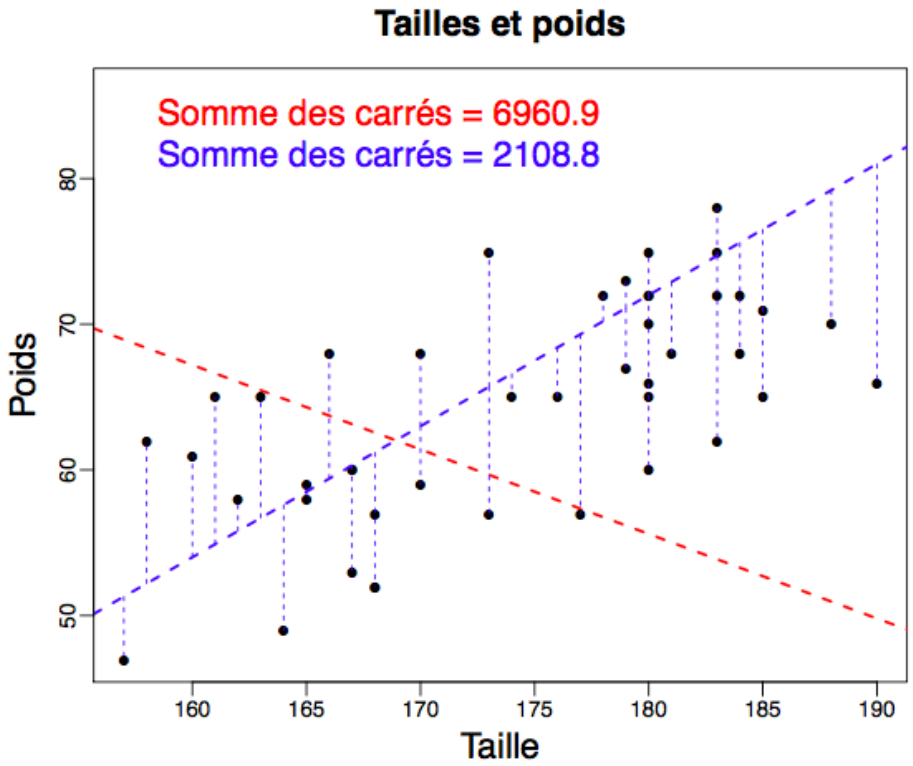
pour quelques autres β_0 et β_1.



On cherche $\widehat{\beta}_0$ et $\widehat{\beta}_1$ telles que

$$\sum_{i=1}^{n}(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2 \leq \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

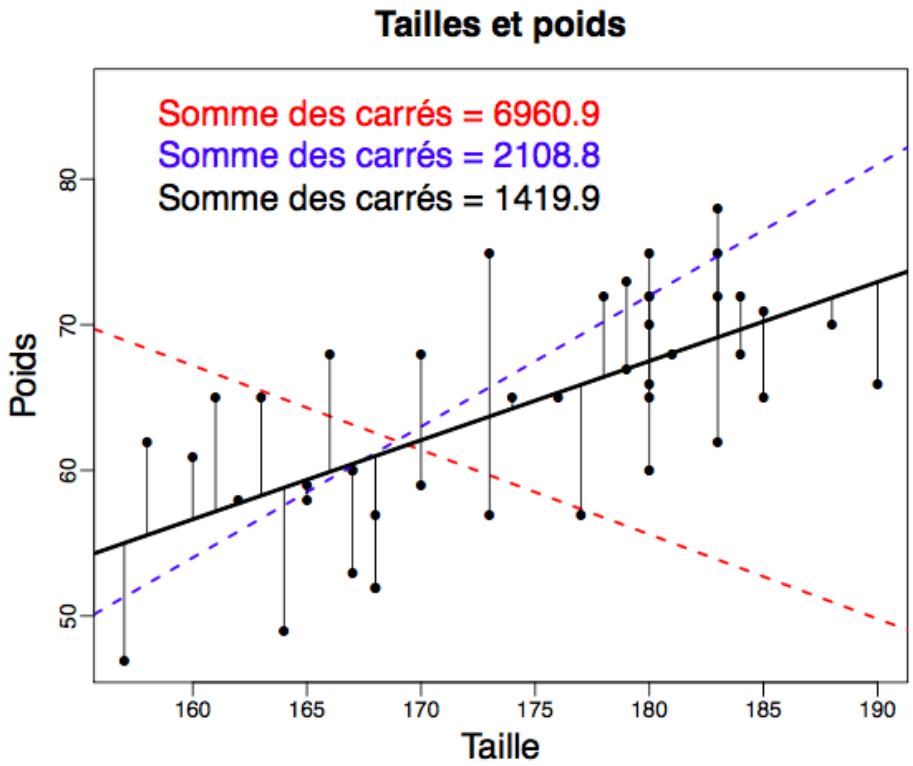
pour quelques autres β_0 et β_1.



On cherche $\widehat{\beta}_0$ et $\widehat{\beta}_1$ telles que

$$\sum_{i=1}^{n}(y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i))^2 \leq \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

pour quelques autres β_0 et β_1.



Pour la régression d'une variable Y par rapport à une variable X, l'ordonnée à l'origine (ou intercept) $\widehat{\beta}_0$ et la pente $\widehat{\beta}_1$ de la droite des moindres carrés peuvent être calculés analytiquement et sont donnés par les formules suivantes:

$$\widehat{\beta}_1 = r(X, Y)\,\frac{s(Y)}{s(X)}$$

$$\widehat{\beta}_0 = m(Y) - \widehat{\beta}_1\, m(X)$$

$\widehat{\beta}_1$ est le coefficient de régression associé à la variable X et quantifie l'association entre X et Y. On voit qu'il est égal à la corrélation multipliée pas le rapport des écarts types de Y et X. Il dépend donc des unités de Y et de X.

Cas des tailles et des poids:

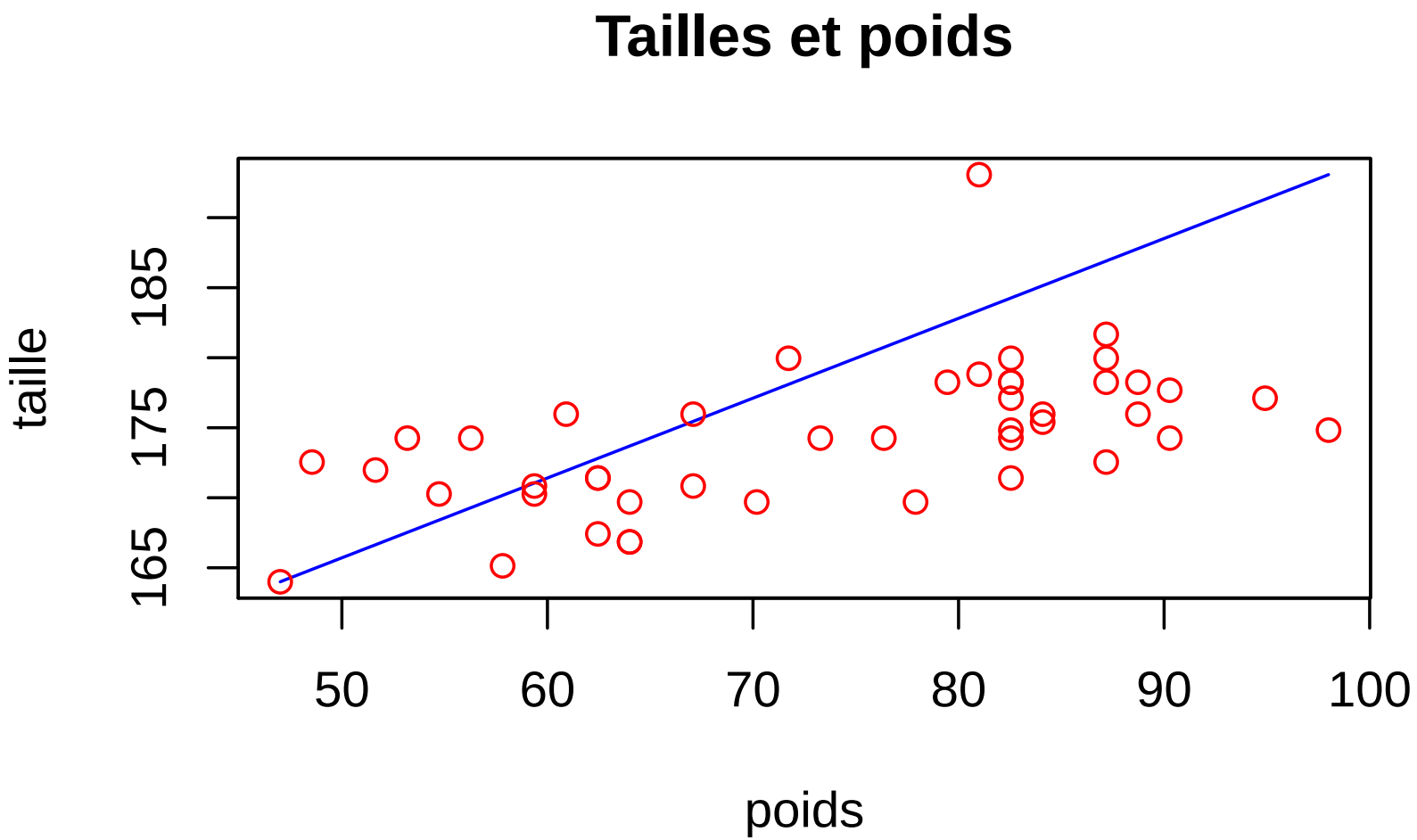
```
beta_1 ← cor(taille,poids)*sd(taille)/sd(poids)
beta_0 ← mean(taille)-beta_1*mean(poids)
beta_1
```

```
[1] 0.5699995
```

```
beta_0
```

```
[1] 137.2105
```

Cas des tailles et des poids: $T = 0.5699995\,P + 137.2104772$



$\widehat{\beta}_1 = 0.5699995$ cm/kg et $\widehat{\beta}_0 = 137.2104772$ cm

Exemple

Une personne entre dans la salle avec un poid de $P = 70$ kg Quelle est sa taille?

$$T = 0.5699995 \times 70 + 137.2104772 = 177.110442 \text{ cm.}$$

Cas des tailles et des poids standarisées:

```
taille_s ← (taille-mean(taille))/sd(taille)
poids_s ← (poids-mean(poids))/sd(poids)
beta_1 ← cor(taille_s,poids_s)*sd(taille_s)/sd(poids_s)
beta_0 ← mean(taille_s)-beta_1*mean(poids_s)
beta_1
```

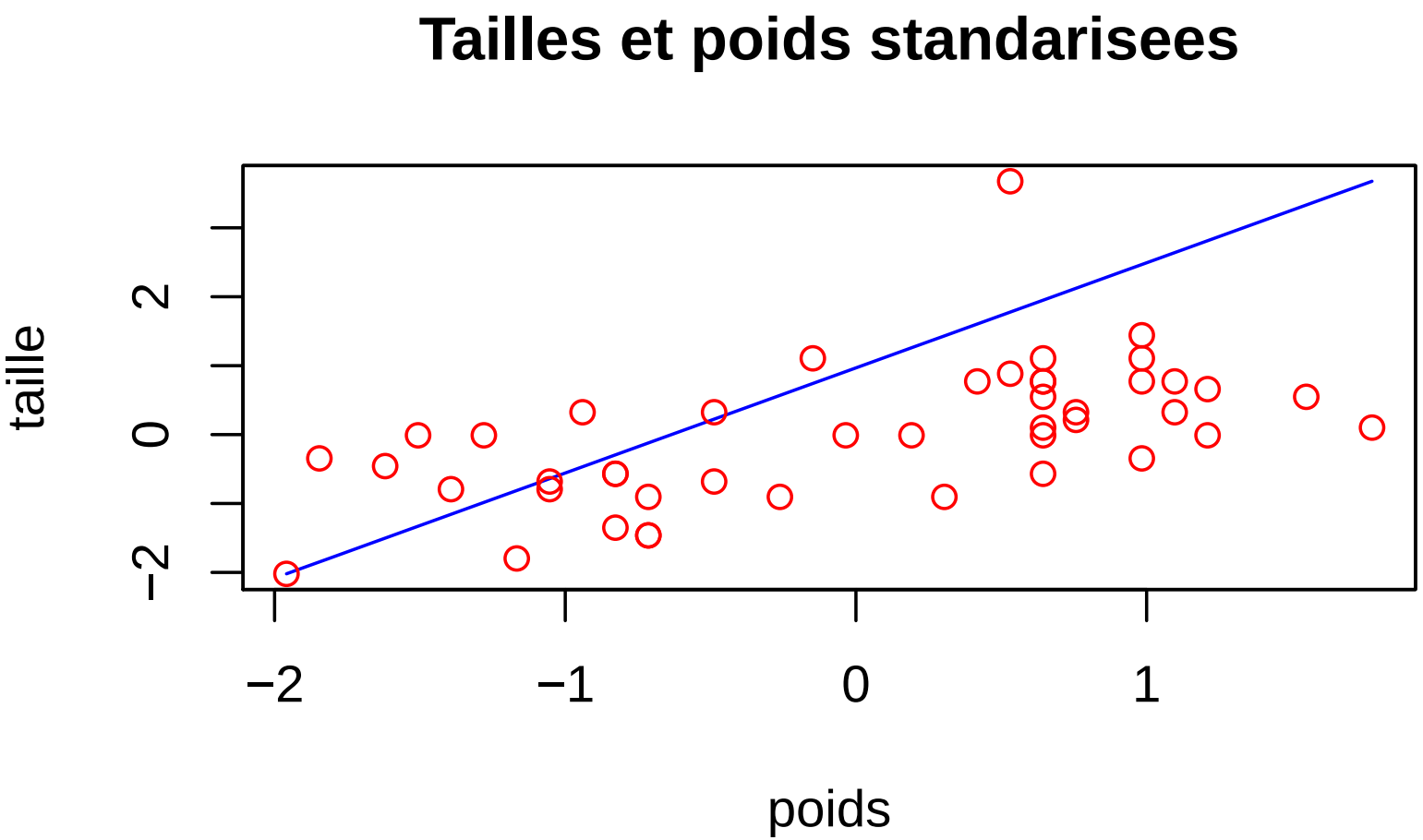
```
[1] 0.5775808
```

```
beta_0
```

```
[1] 9.243326e-16
```

Cas des tailles et des poids standarisées:

$T_s = 0.5775808\, P_s + 9.2433258 \times 10^{-16} \approx 0.5775808\, P_s.$



$\widehat{\beta}_1 = 0.5775808$ et $\widehat{\beta}_0 = 9.2433258 \times 10^{-16}$

Exemple

Une personne entre dans la salle avec un poid de $P = 70$ kg Quelle est sa taille?

$$P_s = \frac{P - 65.0888889}{8.9539789}.$$

$$T_s = 0.5775808 \times \frac{70 - 65.0888889}{8.9539789} + 9.2433258 \times 10^{-16} = 0.3167936$$

Alors, comme $T = s(T) \times T_s + m(T)$ on a

$$T = 8.8364489 \times 0.3167936 + 174.3111111 = 177.110442 \text{ cm.}$$

Terminologie et définitions:

- Dans le contexte de la régression, on convient d'appeler Y la variable réponse ou la variable dépendante et X la variable explicative ou la variable indépendante.
- Lorsqu'on calcule les valeurs de $\widehat{\beta}_0$ et de $\widehat{\beta}_1$, on fait une estimation d'un modèle sous-jacent que l'on postule au niveau de la population:

$$Y = \beta_1 X + \beta_0 + \varepsilon,$$

ou β_0 et β_1 sont les vraies valeurs de l'intercept et de la pente au niveau de la population et ε est une variable appelée l'erreur. En statistique, on utilise souvent le (chapeau) $\widehat{\text{pour}}$ indiquer qu'une variable est une estimation d'un paramètre.

- $\widehat{Y} = \widehat{\beta}_1 X + \widehat{\beta}_0$ est la variable des réponses calculées ($\widehat{y}_i$ est la valeur sur la droite correspondant à x_i).
- $\varepsilon = \widehat{Y} - Y$ est la variable des résidus

Propriétés

- $$\begin{array}{ccccc} Y & = & \widehat{Y} & + & \widehat{\varepsilon} \\ \text{réponse observée} & = & \text{réponse calculée} & + & \text{résidu} \end{array}$$
- La droite des moindres carrés passe par le point $(m(X), m(Y))$.
- La somme des résidus est nule:

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = \sum_{i=1}^{n} (y_i - (\widehat{\beta}_1 x_i + \widehat{\beta}_0)) = 0.$$

Différence entre corrélation et coefficient de régression

Corrélation et cofficient de régression sont deux mesures complémentaires de l'association entre deux variables. Ils renseignent chacun sur un aspect différent de la relation entre les deux variables: la corrélation informe sur la précision avec laquelle on peut prédire l'une à partir de l'autre, tandis que le coefficient de régression mesure l'importance de l'"effet" moyen de l'une sur l'autre.

On considère dans la base des donées `nutriage.csv` les variables `taille` et `poids` pour étudier l'association linéaire entre les deux variables.

```
donnees<-read.csv(
"https://afalco.000webhostapp.com/cursos/nutriage.csv"
,header=TRUE)
```

On considère la variable `taille.chapeau` définie comme:

$$\texttt{taille.chapeau} = \beta_0 + \beta_1 \, \texttt{poids}$$

en R:

```
beta0 <- 100
beta1 <-  0.8
taille.chapeau <- beta0 + beta1*poids
```

On compare les variables `taille` et `taille.chapeu` avec

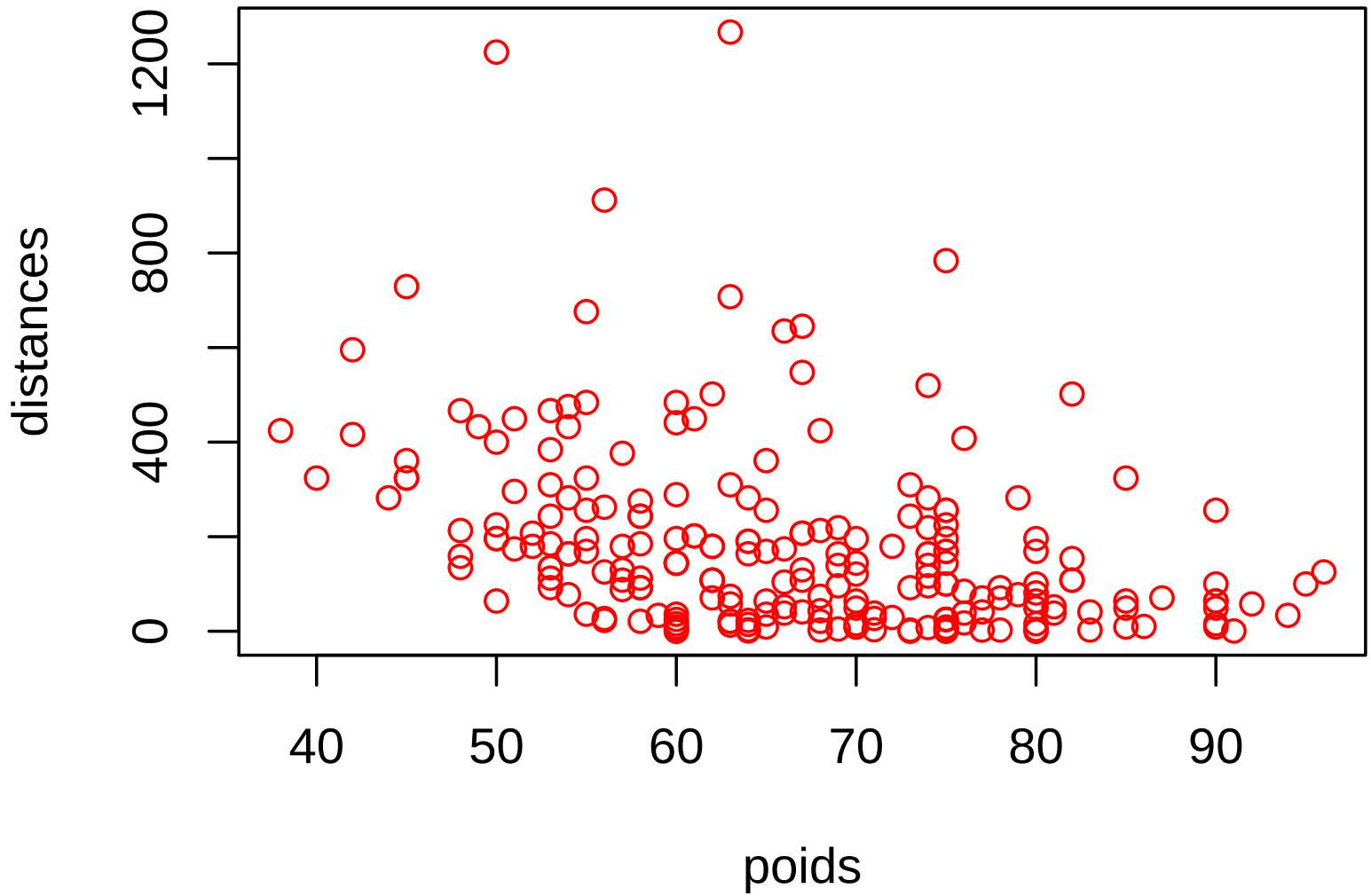
$$\varepsilon = \texttt{taille.chapeau} - \texttt{taille}$$

et

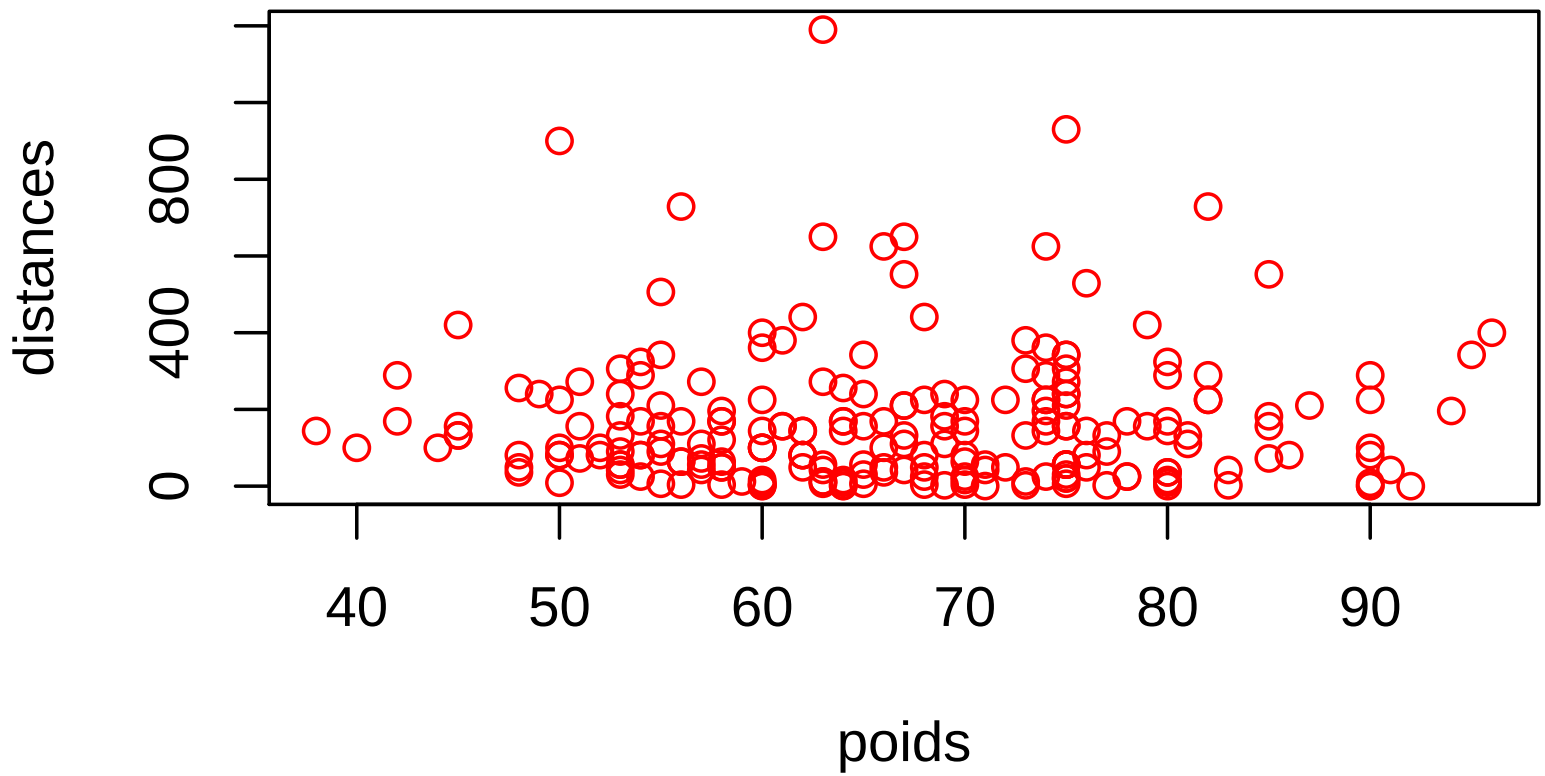
$$\varepsilon^2 = (\texttt{taille.chapeau} - \texttt{taille})^2$$

qu'on appelle `distances`: en R

```
residus ← taille.chapeau-taille
distances ← residus^2
```


distances
0
400
800
1200
40
50
60
70
80
90
poids

```
beta0 ← 120
beta1 ←  0.5
```



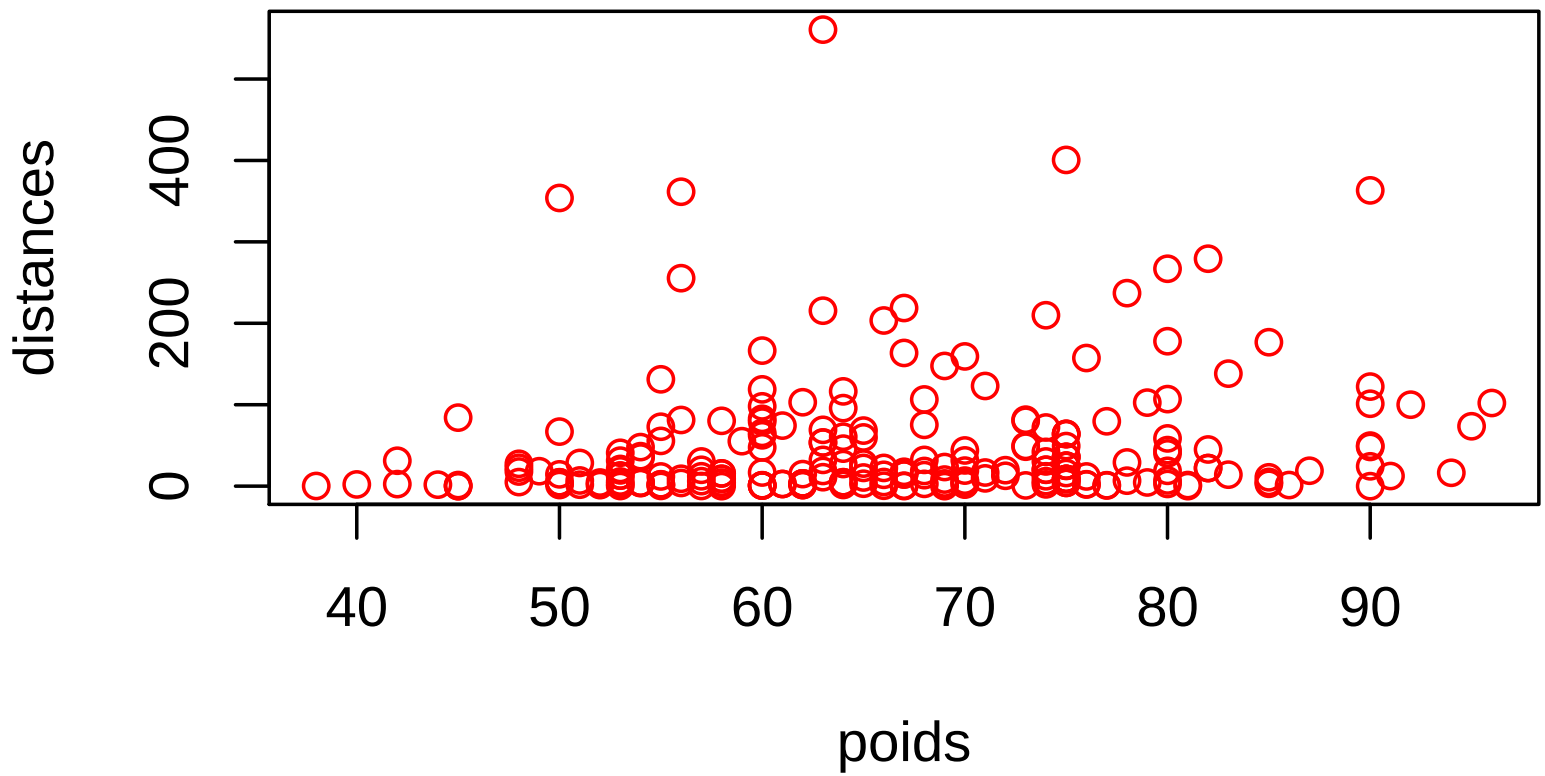
Pour obtenir des valeurs optimales:

```
lm(taille~poids)
```

```
Call:
lm(formula = taille ~ poids)

Coefficients:
(Intercept)        poids
   132.5900       0.4719
```

```
beta0 ← 132.5900
beta1 ← 0.4719
```



Calcule explicite des valeurs optimales β_0 et β_1 notées: $\widehat{\beta}_0$ et $\widehat{\beta}_1$.

```
beta1.chapeau <- cor(poids,taille)*sd(taille)/sd(poids)
beta0.chapeau <- mean(taille)-beta1.chapeau*mean(poids)
beta0.chapeau
```

```
[1] 132.59
```

```
beta1.chapeau
```

```
[1] 0.4718581
```

Statistique des erreurs d'approximation:

```
mean(residus)
```

```
[1] 0.002820796
```

```
sd(residus)
```

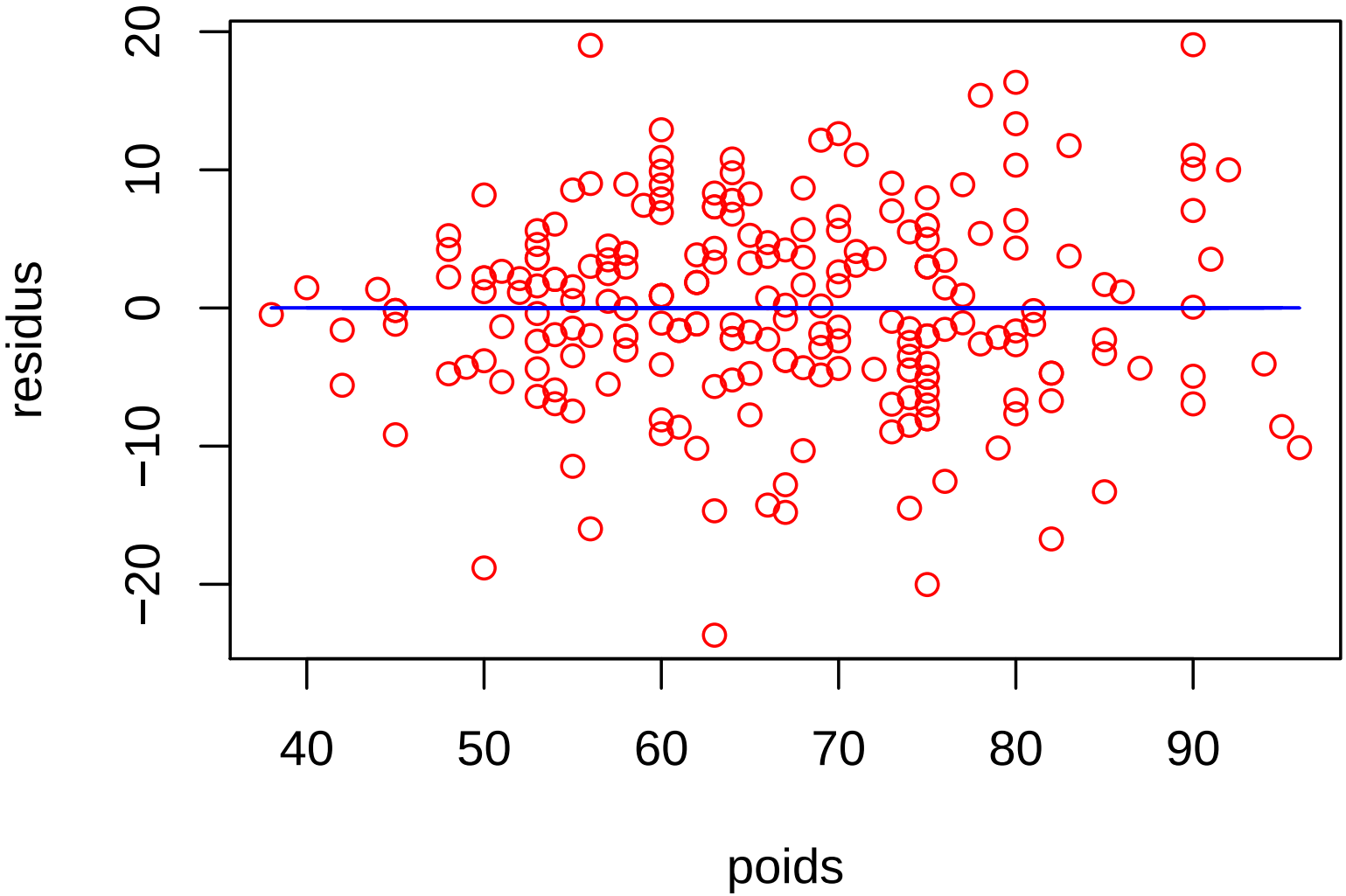
```
[1] 6.987165
```

```
mean(distances)
```

```
[1] 48.60446
```

```
erreur.optimal ← sqrt(mean(distances))
erreur.optimal
```

```
[1] 6.97169
```

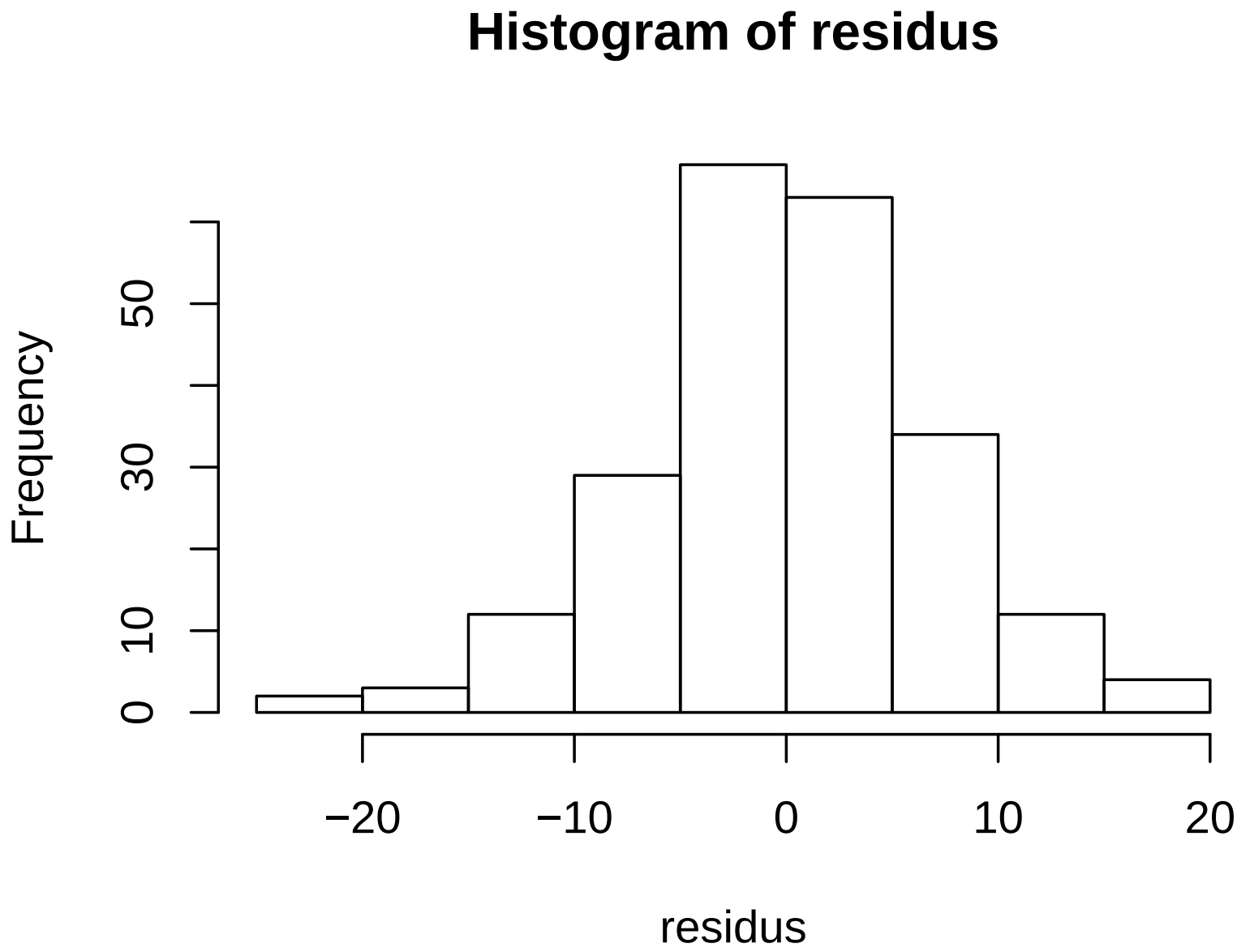

residus
20
10
0
−10
−20
40
50
60
70
80
90
poids

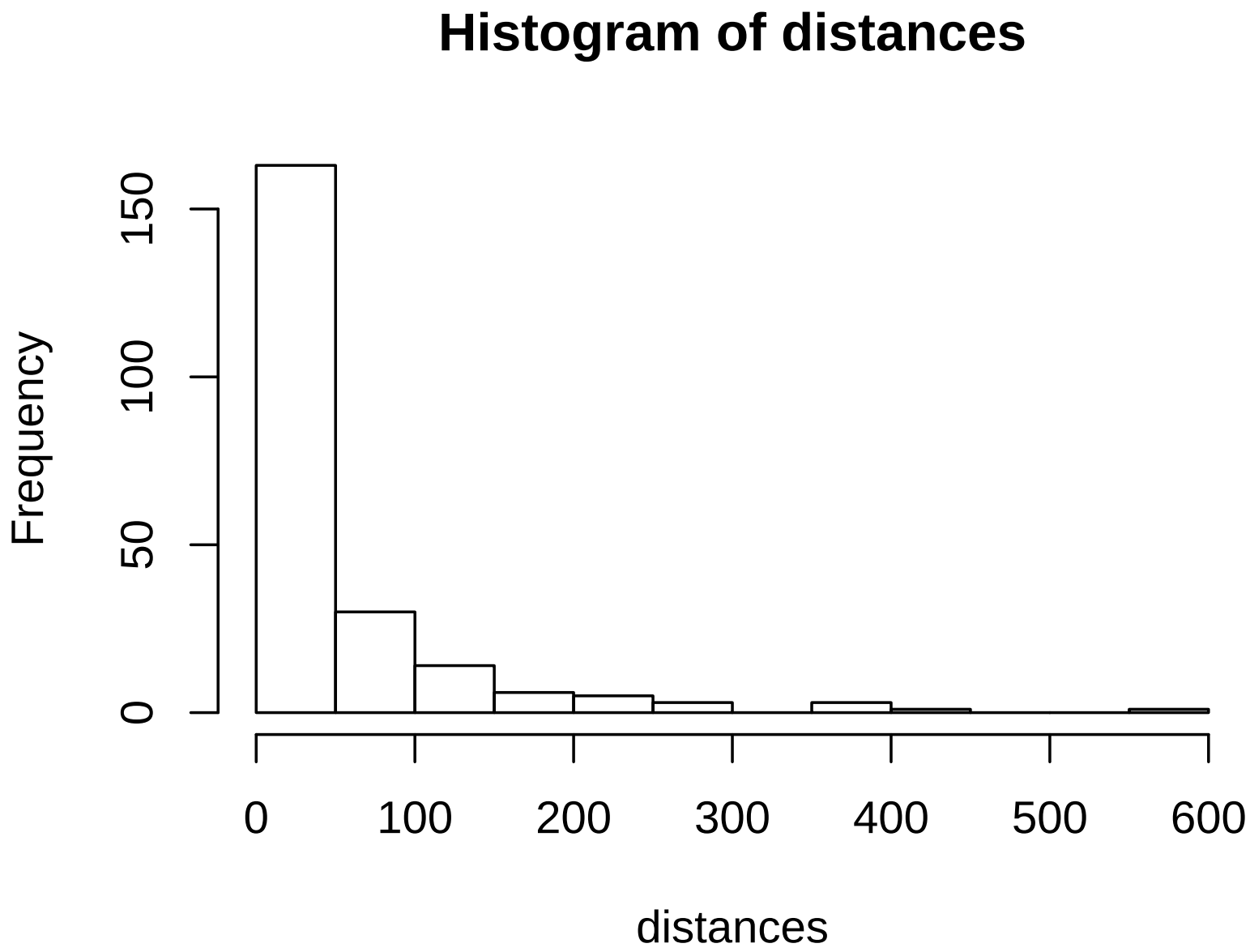
Si $m(\varepsilon) = m(taille.chapeau - taille) \approx 0$ alors la variance des résidus:

$$s^2(\varepsilon) = m((\varepsilon - m(\varepsilon))^2) \approx m(\varepsilon^2) = \texttt{distances}.$$

```
summary(residus)
```

```
      Min.    1st Qu.     Median       Mean    3rd Qu.
Max.
-23.680300  -4.371425   0.010600   0.002821   4.232725  19.061000
```


Histogram of residus
Frequency
0
10
30
50
−20
−10
0
10
20
residus


Histogram of distances
Frequency
0
50
100
150
0
100
200
300
400
500
600
distances