# Corrélation et régression

## Problème

- Peut-on utiliser un mètre ruban pour peser un ours ?

## Introduction

- Des chercheurs ont étudié des ours en les anesthésiant pour mesurer des facteurs tels que l'âge,le sexe, la taille et le poids. Le poids des ours étant très élevé, il est difficile de les soulever et de les peser sur le terrain.
- Est –il possible de déterminer le poids d'un ours à partir d'autres type de mesures plus faciles à obtenir ?

Les données suivantes montrent –elles l'existence d'une association entre la taille d'un ours et son poids ? si oui, de quelle type est cette association ? si un chercheur anesthésie un ours et le mesure à 180cm, comment utiliser cette taille pour prédire son poids ? ces questions sont abordées dans ce chapitre.

Tableau 1 : « taille et poids d'ours mâles »

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| x taille (cm) | 134.6 | 171.5 | 182.9 | 182.9 | 186.7 | 174 | 185.4 | 94 |
| y poids (kg) | 36.3 | 156 | 188.7 | 157.9 | 118.8 | 163.3 | 150.6 | 15.4 |

# I. Corrélation

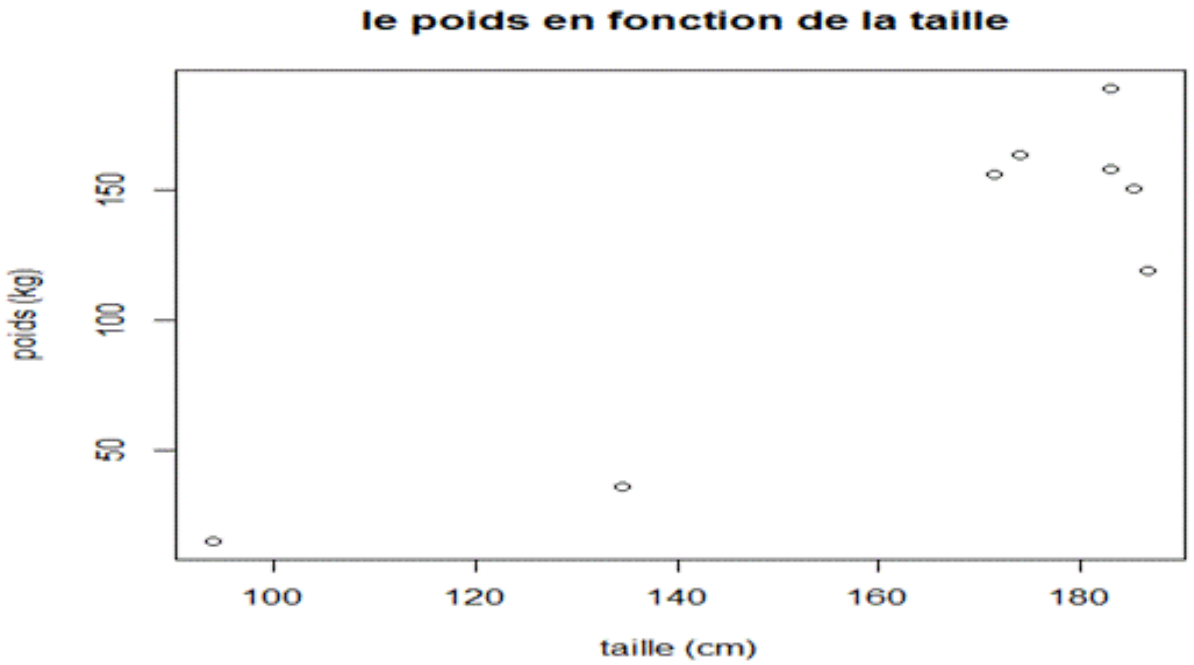
- L'objectif principal de cette section est d'analyser un ensemble de données en paires et de déterminer s'il semble y avoir une association entre les deux variables. En statistique, une telle relation est appelée corrélation.

## Exploration des données

- Avant d'effectuer les calculs de la corrélation entre les deux variables, il est judicieux de commencer par explorer les données au moyen d'un diagramme de dispersion.

# Diagramme de dispersion

- C’est un graphique dans lequel les données (x, y) sont placées selon un axe horizontal (x) et un axe vertical (y). chaque paire individuelle (x,y) est représentée par un point.
- Les données du tableau sont représentées par le diagramme suivant



Lors de l'examen d'un tel diagramme, il est important de prêter attention à la forme générale du nuage des points. La figure indique que les tailles les plus élevées semblent associées aux poids les plus importants. Cela suggère une association entre taille et poids chez les ours.

## Coefficient de corrélation linéaire

- L'examen visuel des digrammes de dispersion étant très subjectif, nous avons besoin d'une mesure objective. Nous utilisons pour cela le coefficient de corrélation linéaire r.

Le coefficient de corrélation linéaire r mesure l'intensité de l'association linéaire entre les valeurs de x et y liées et issues d'un échantillon. Cette valeur est calculée en utilisant la formule suivante :

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

# Exemple « taille et poids des ours »

A partir des données du tableau 1 trouver la valeur du coefficient linéaire r.

$\sum x = 1312$

$\sum y = 987$

$\sum x^2 = 222776.3$

$\sum y^2 = 149891.6$

$\sum xy = 174996.1$

r= 0.897

## Propriétés du coefficient de corrélation r

- $-1 \leq r \leq +1$
- r mesure l'intensité d'une association linéaire.

## Interprétation de r

- Conclure qu'il y a une corrélation linéaire significative entre x et y implique qu'il est possible de trouver une équation linéaire qui exprime y en fonction de x et que cette équation peut être utilisée pour prédire les valeurs de y pour des valeurs de x données.

  La valeur de $r^2$ est la proportion de la variation de y expliquée par l'association linéaire entre x et y ($r^2$ est appelé coefficient de détermination)

# Exemple

- A partir de données du tableau 1, nous avons trouvé que le coefficient de corrélation linéaire était r=0.897. quelle proportion de la variation du poids des ours peut être expliquée par la relation linéaire entre le poids et la taille des ours ?

Solution avec r=0.897 on obtient $r^2$=0.805

# Interprétation

- nous pouvons conclure que 80.5% (environ 81%) de la variation du poids des ours peut être expliquée par l'association entre le poids des ours et leur taille. Cela implique que 19% de la variation du poids des ours doit être expliquée par d'autres facteurs que leurs taille.

## Erreurs communes au sujet de la corrélation

- Voici les trois sources d'erreurs les plus communes dans l'interprétation des résultats concernant la corrélation:
  - une source d'erreur courante est de considérer que la corrélation implique causalité.( nous ne pouvons pas affirmer que de plus grandes tailles causent des poids plus importants. Le poids des ours peut être affectés par d'autres variables, dites cachées).

1- Une autre erreur courante se produit avec des données basées sur des moyennes. Les moyennes éliminent la dispersion individuelle et leur utilisation peut « gonfler »le coefficient de corrélation.

1- Une troisième source d'erreur est liée à la propriété de linéarité. Une association peut exister entre x et y même si la corrélation linéaire n'est pas significative.

# Test d'hypothèse de la corrélation

ρ le coefficient de corrélation de la population

$H_0$ : ρ=0 (il n'y a pas de corrélation linéaire significative)

$H_1$ : ρ≠ 0 (corrélation linéaire significative)

$H_1$ : ρ>0 (corrélation linéaire positive)

$H_1$ : ρ<0 (corrélation linéaire négative)

Statistique de test

$$t=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

valeurs critiques utiliser la table de student avec n-2 degré de liberté.

# Exemple « taille et poids des ours »

- En utilisant le tableau 1, tester l'affirmation qu'il y a une corrélation linéaire significative entre la taille et le poids des ours.

## Solution

Un exemple précédent montre que les conditions requises sont respectées.

Affirmer qu'il y a une corrélation linéaire significative est équivalent à dire que $\rho \neq 0$

Nous testons les hypothèses suivantes

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.897}{\sqrt{\frac{1-0.897^2}{8-2}}}$$

$t = 4.971$

il s'agit d'un test bilatéral alors il y a deux valeurs critiques.+2.447 et -2.447

(la table de student la colonne 0.05 et la ligne n-2=6)

$t > 2.447$ le rejet de $H_0$.

## Conclusion

- il y a donc suffisamment d'éléments pour confirmer l'hypothèse d'une corrélation linéaire entre la taille des ours et leur poids.

# II. Régression

- La section précédente montrait comment analyser des paires de données pour déterminer la présence d'une corrélation linéaire significative entre deux variables.
- L'objectif de la présente section est de décrire l'association entre deux variables en établissant l'équation et le graphique de la droite qui représente cette relation.

Cette droite est appelée droite de régression et son équation est l'équation de régression.

Cette équation exprime une association entre x (appelé variable indépendante, variable explicative ou variable prédictive) et $\hat{y}$ (appelée variable dépendante ou variable réponse), elle est exprimée sous la forme

$$\hat{y} = b_0 + b_1 x$$

$b_0$ est l'ordonnée à l'origine

$b_1$ est la pente.

Nous utiliserons des paires de données pour estimer l'équation de régression.

# Notation pour la droite de régression

| Paramètre (population) | Statistique (échantillon) |
|---|---|
| Ordonnée à l'origine $\beta_0$ | $b_0$ |
| Pente de la droite $\beta_1$ | $b_1$ |
| Equation de la droite $y=\beta_0 + \beta_1 x$ | $\hat{y} = b_0 + b_1 x$ |

$$b_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

# Exemple « taille et poids des ours »

- A partir des données du tableau 1, nous avons vu que r= 0.897. En utilisant les mêmes données, calculer l'équation de la droite de régression.

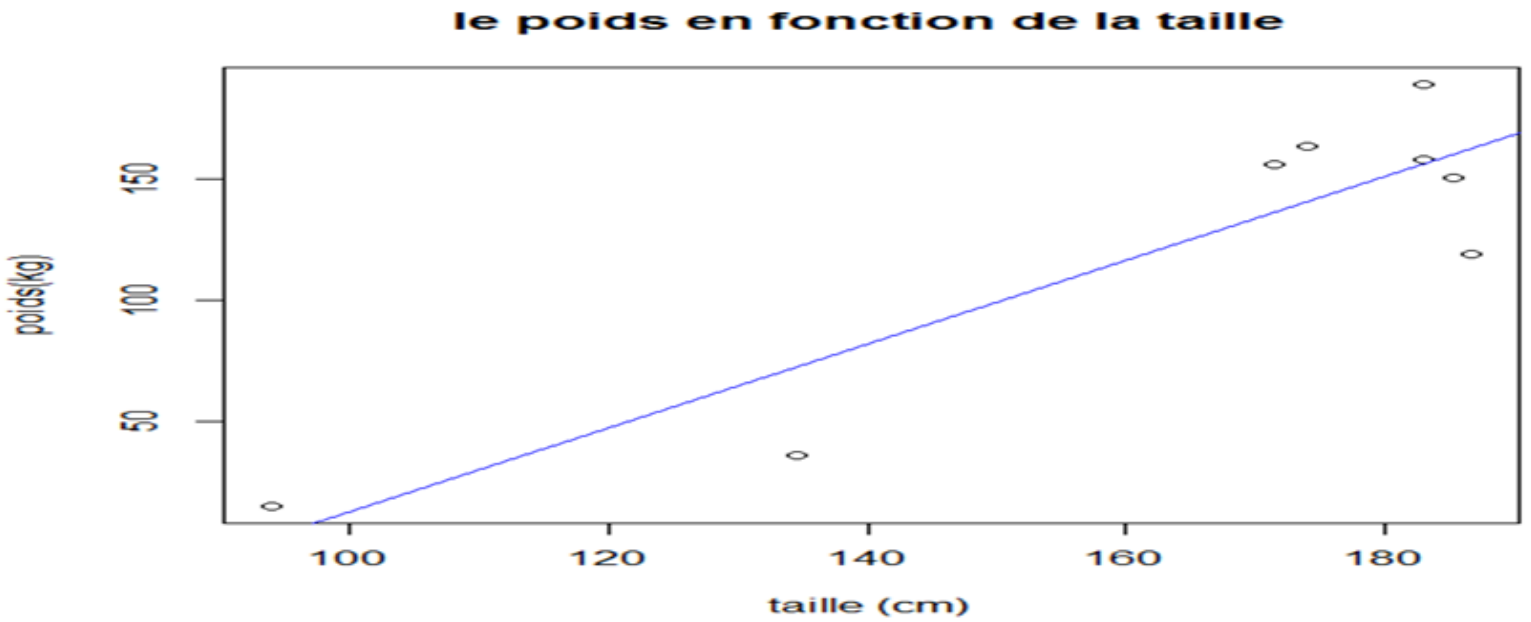
## Solution

La forme du nuage de points suggère une relation. Il n'y a pas de valeurs extrêmes et nous considérons que l'échantillon a été sélectionné aléatoirement.

D'après les formules précédentes, on a

$b_0$=-160 et $b_1$=1.73. cela fournit l'équation de régression :

$\hat{y}$=-160+1.73x




## Utilisation de l'équation de régression

- Les équations de régression peuvent servir à prédire les valeurs d'une variable à partir des valeurs données d'une autre variable.
- Si la droite de régression s'ajuste bien aux données, il est possible d'utiliser son équation pour faire des prédictions, à condition que celles-ci ne dépassent pas la gamme des valeurs de l'échantillon.

Par exemple, à partir du tableau 1 comme la droite de régression s'ajuste bien aux données tailles/poids des ours, si on mesure la taille d'un ours à 180cm, nous pouvons prédire son poids en substituant x=180 dans l'équation de régression $\hat{y}$=-160+1.73x

Les résultats suivants montrent que si un ours mesure 180cm son poids peut être prédit à 151kg

$\hat{y}$=-160+1.73x

-160+1.73(180) = 151

# Variation expliquée et non expliquée

- La section précédente introduit le concept de corrélation et l'utilisation du coefficient de corrélation r pour déterminer l'existence d'une relation significative entre deux variables x et y.
- La valeur de r donne également des informations sur la dispersion de points autour de la droite de régression.

Supposons un grand échantillon de données pour deux variables x et y correspondants aux résultats suivants :

- Il y a une corrélation linéaire significative
- L'équation de la droite de régression est $\hat{y}$ = 3+2x
- La moyenne des valeurs de y est $\bar{y}$ = 9
- Une des paires de données correspond à x=5 et y=19
- Le point (5, 13) est un des points de la droite de régression car substituer x=5 dans l'équation de régression donne $\hat{y}$ =13
- Le point (5,13) se situe sur la droite, mais le point (5,19) se situe l'écart de cette droite.
- Déviation totale de (5,19) =y- $\bar{y}$
  =19-9=10
- Déviation expliquée de (5,19)= $\hat{y}$- $\bar{y}$
  = 13-9=4
- Déviation non expliquée de (5,1)= y- $\hat{y}$ =19-13=6.

La déviation totale (par rapport à la moyenne) du point (x,y) est la distance verticale y- $\bar{y}$

La déviation expliquée est la distance verticale

$\hat{y}$- $\bar{y}$

La déviation non expliquée est la distance verticale y- $\hat{y}$

déviation totale= déviation expliquée+ déviation non expliquée.

Ou

$$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

# Coefficient de détermination

C'est la qualité de variation de y qui est expliquée par la droite de régression.

$$r^2 = \frac{variation \quad expliqu\ ée}{variation \quad totale}$$

## Exemple « cholestérol et poids ».

- Le coefficient de corrélation linéaire entre le taux de cholestérol et l'indice de masse corporelle dans un échantillon de 40 femmes est r=0.482.
- Trouver le pourcentage de variation de y (IMC) expliqué par l'association linéaire entre le taux de cholestérol est l'IMC.

# Solution

Le coefficient de détermination $r^2=0.482^2 = 0.232$, indiquant que le rapport entre la variation expliquée et la variation totale de y est 0.232.

Nous pouvons donc dire que 23.2% de la variation totale de y est expliquée par l'équation de régression.

En termes biologiques, 23.2% de la variation totale de l'IMC peut être expliquée par la variation du taux de cholestérol ; les 76.8% restants sont dus à d'autres facteurs, comme le poids, la taille ou des facteurs génétiques.

## Valeurs extrêmes et valeurs influentes

- Une analyse de régression ou de corrélation de données bivariées devrait considérer l'effet des valeurs extrêmes et des valeurs influentes, définies ci-dessous
- Dans un diagramme de dispersion une valeur extrême correspond à un point qui se situe très à l'écart des autres.

- Les données peuvent également inclure une ou plusieurs valeurs influentes, qui sont les points qui contribuent de façon importante à la droite. Pour l'identifier, on établit la droite de régression avec et sans ce point. Si la variation produite est importante, le point est influent.