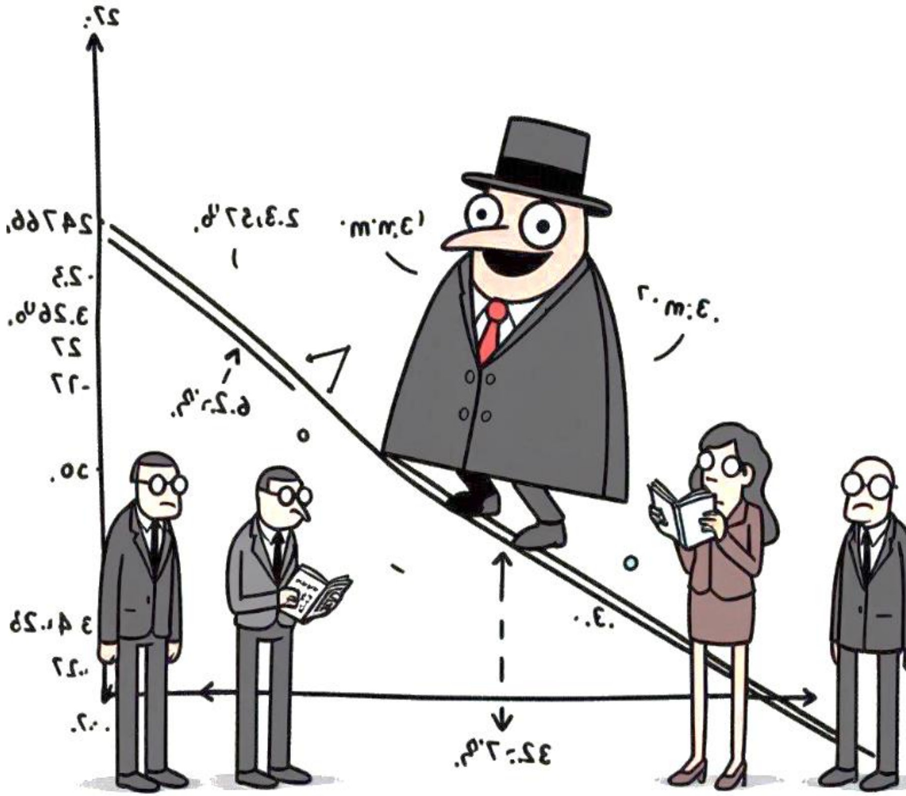
STATISTIQUES POUR MEDECINS

Régression linéaire

Christophe Combescure

Unité d'Appui Méthodologique du CRC



Hôpitaux Universitaires Genève

FACULTÉ DE MÉDECINE

UNIVERSITÉ DE GENÈVE

Rappel: Tests statistiques

- Procédure pour choisir entre hypothèse nulle et hypothèse alterne
- D’abord on formule les hypothèses
- H_0: pas d’association
  - survie identique avec le traitement expérimental et le comparateur
  - aucune différence de taille entre femmes et hommes
- H_A: association est présente
- Ensuite on récolte les données
  - observations de survie chez des patients traités et non-traités
  - mesure des tailles d’hommes et de femmes
- On fait le test, dont le résultat est significatif ou non-significatif

Rappel: Tests statistiques

- Si l'hypothèse nulle est vraie:
  - résultat non-significatif (réponse correcte) dans 95% des cas
  - résultat significatif (erreur de type 1) dans 5% des cas
- Erreur de type 2:
  - résultat non-significatif lorsque l'hypothèse alterne est vraie
  - risque contrôlé à 20% (ou 10%) par un calcul de la taille d'échantillon
  - puissance = 100% - risque erreur de type 2
- Si résultat significatif:
  - soit l'hypothèse alterne est vraie
  - soit l'hypothèse nulle est vraie et on a fait une erreur de type 1

Option privilégiée dans l'interprétation

- Si résultat non-significatif:
  - soit l'hypothèse nulle est vraie
  - soit l'hypothèse alterne est vraie on a fait une erreur de type 2

Rappel: Valeur p

- Mesure à quel point les résultats observés contredisent l'hypothèse nulle
- Plus p est petit, plus les résultats contredisent H_0
- Valeur p: probabilité de survenue de la différence observée ou d'une différence plus importante si l'hypothèse nulle était vraie
- $p \leq 0.05$ équivaut à un résultat du test significatif
- $p > 0.05$ équivaut à un résultat du test non-significatif

Rappel: Intervalle de confiance

- On fait des hypothèses sur des paramètres (vraies valeurs qui décrivent l'univers), mais on observe des estimations (issus d'échantillons limités)
- Intervalle de confiance (IC): ensemble des valeurs du paramètre qui sont compatibles avec l'estimateur observé
- IC à 95%: si on répétait l'étude un grand nombre de fois, 95% des IC calculés contiendraient la vraie valeur du paramètre (mais dans le cas particulier on ne sait pas si c'est le cas…)
- Plus l'IC est étroit, plus l'estimation est précise
- Si IC contient la valeur du paramètre correspondant à H_0: test non-significatif
- Si IC exclut la valeur du paramètre correspondant à H_0: test significatif

Objectifs

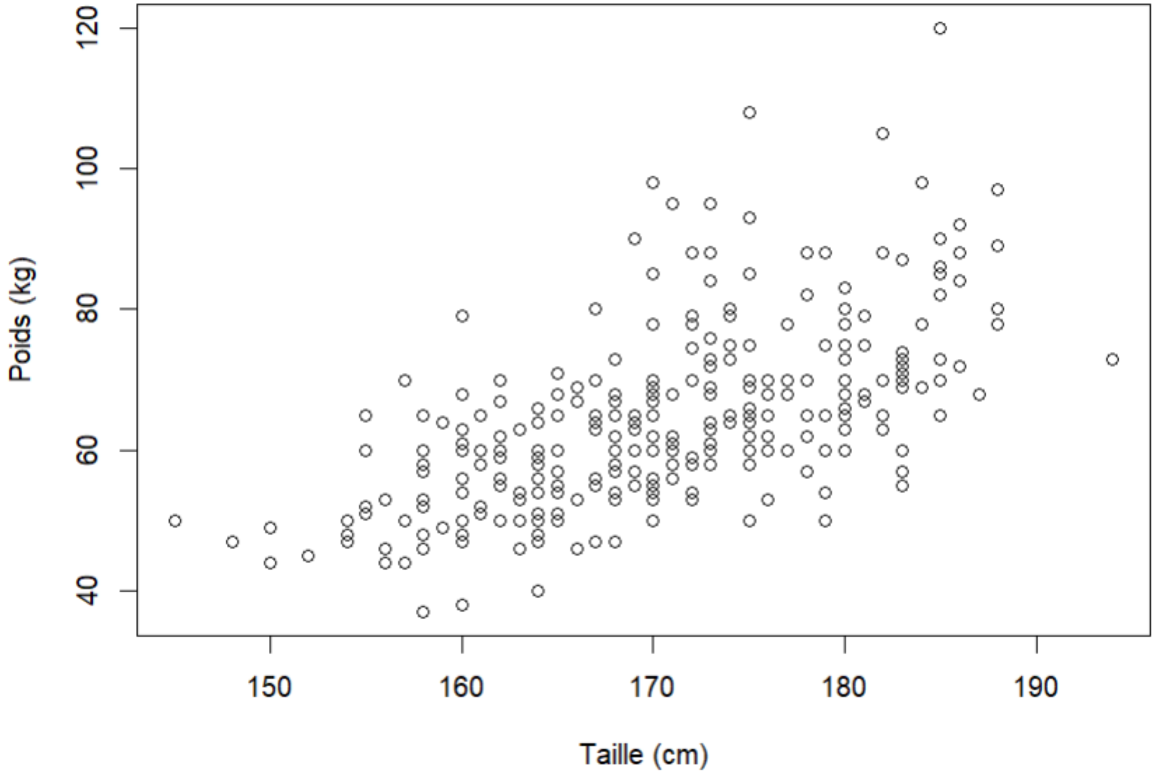
- Comprendre les notions suivantes
  - corrélation entre deux variables continues
  - modélisation
  - régression linéaire simple
  - régression linéaire multiple

Chapitres Petrie/Sabin
26 (corrélation)
27 – 29 (modèle de régression linéaire)

Exemple du questionnaire

Question de recherche:
La taille des individus est-elle liée au poids corporel ?

Relation entre poids et taille



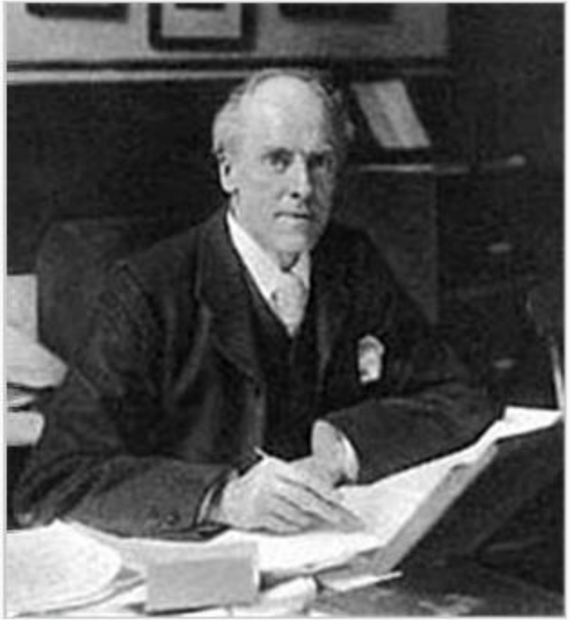
Plus la taille est grande, plus le poids est élevé

Plus le poids est élevé, plus la taille est grande

Corrélation

- En statistique, la corrélation est la relation entre 2 variables (continues et mesurées chez les mêmes sujets)

- Karl Pearson propose en 1896 une formule mathématique pour la notion de corrélation et un estimateur de cette grandeur: **coefficient de corrélation de Pearson**



Karl Pearson

Coefficient de corrélation r

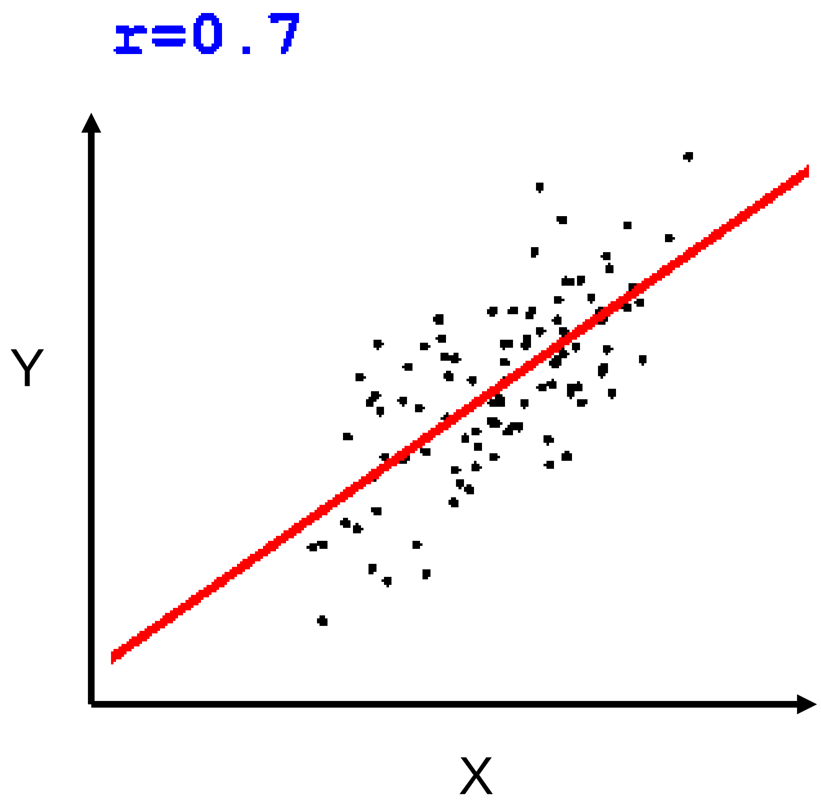
- Formule de la corrélation (notée ***r***) entre X et Y:

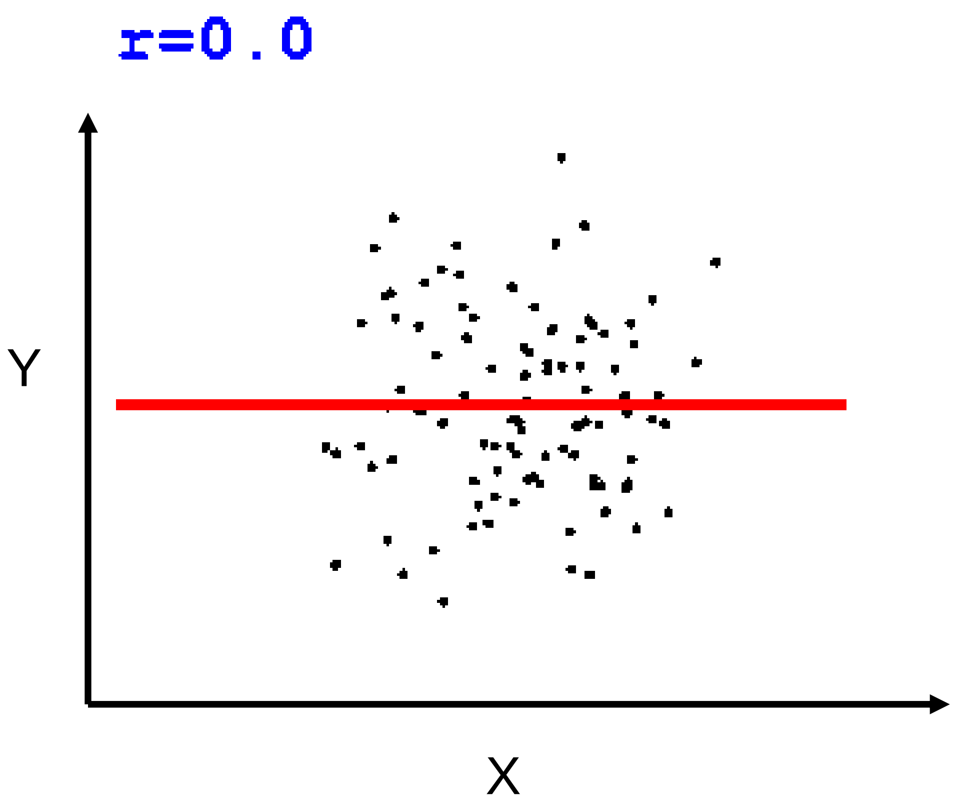
$$r = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{(n-1)s_x s_y}$$

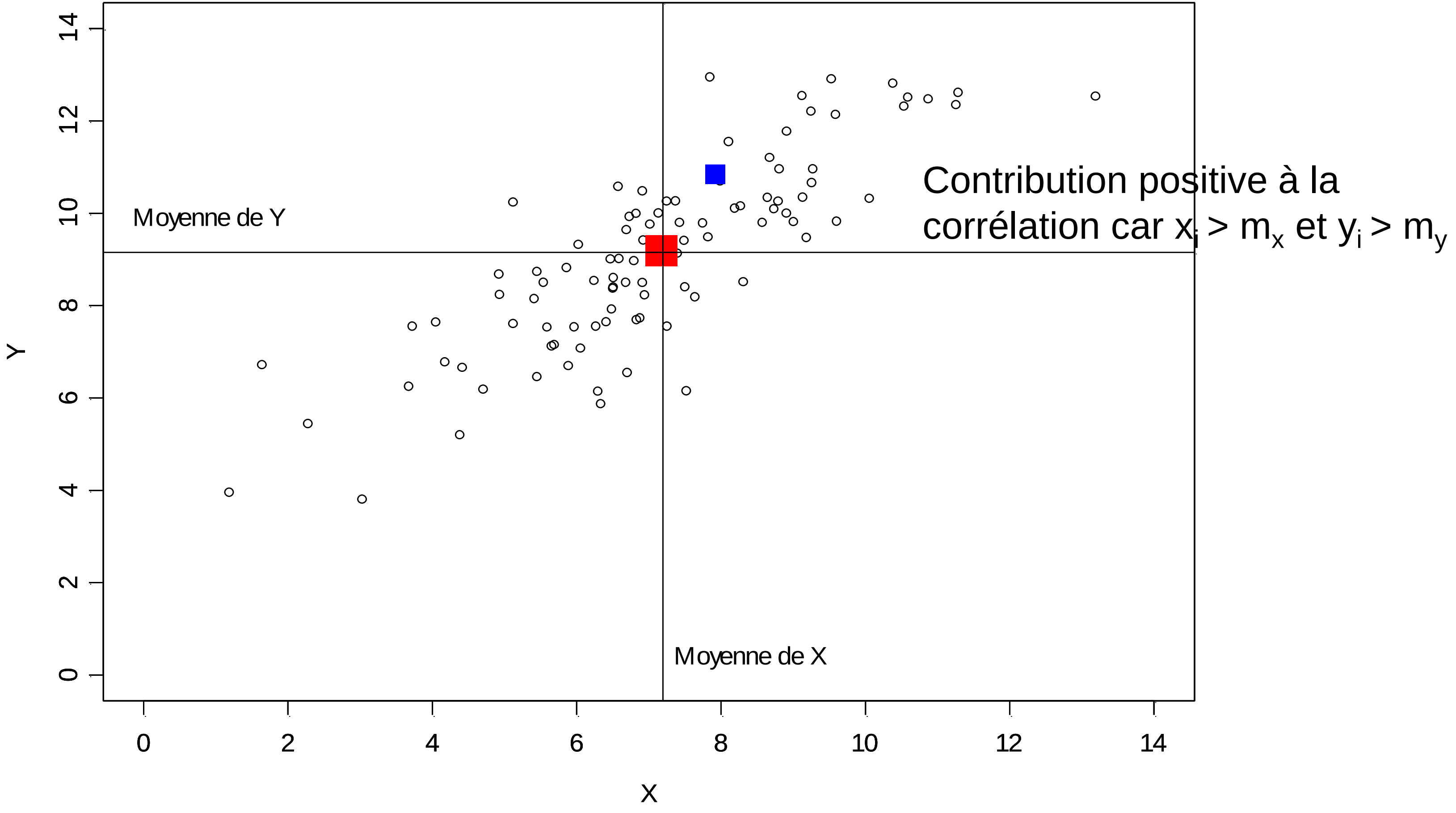
- x_i et y_i sont les mesures des variables X et Y chez le sujet i
- m_x et m_y sont les moyennes de X et de Y dans l'échantillon
- n est le nombre de sujets dans l'échantillon
- s_x et s_y sont les écarts types de X et Y dans l'échantillon

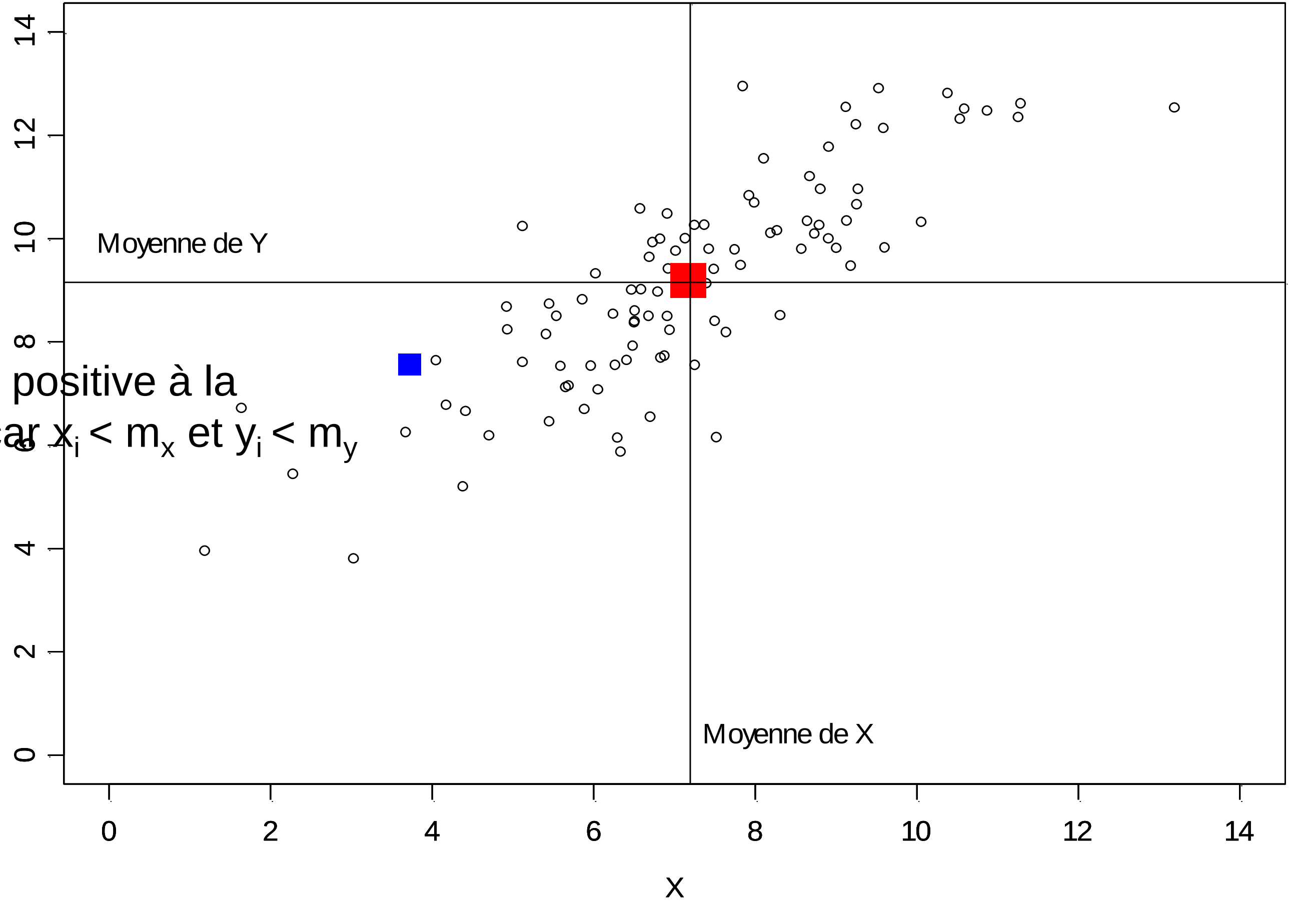
- Le dénominateur contraint ***r*** à être entre -1 et 1

Coefficient de corrélation r

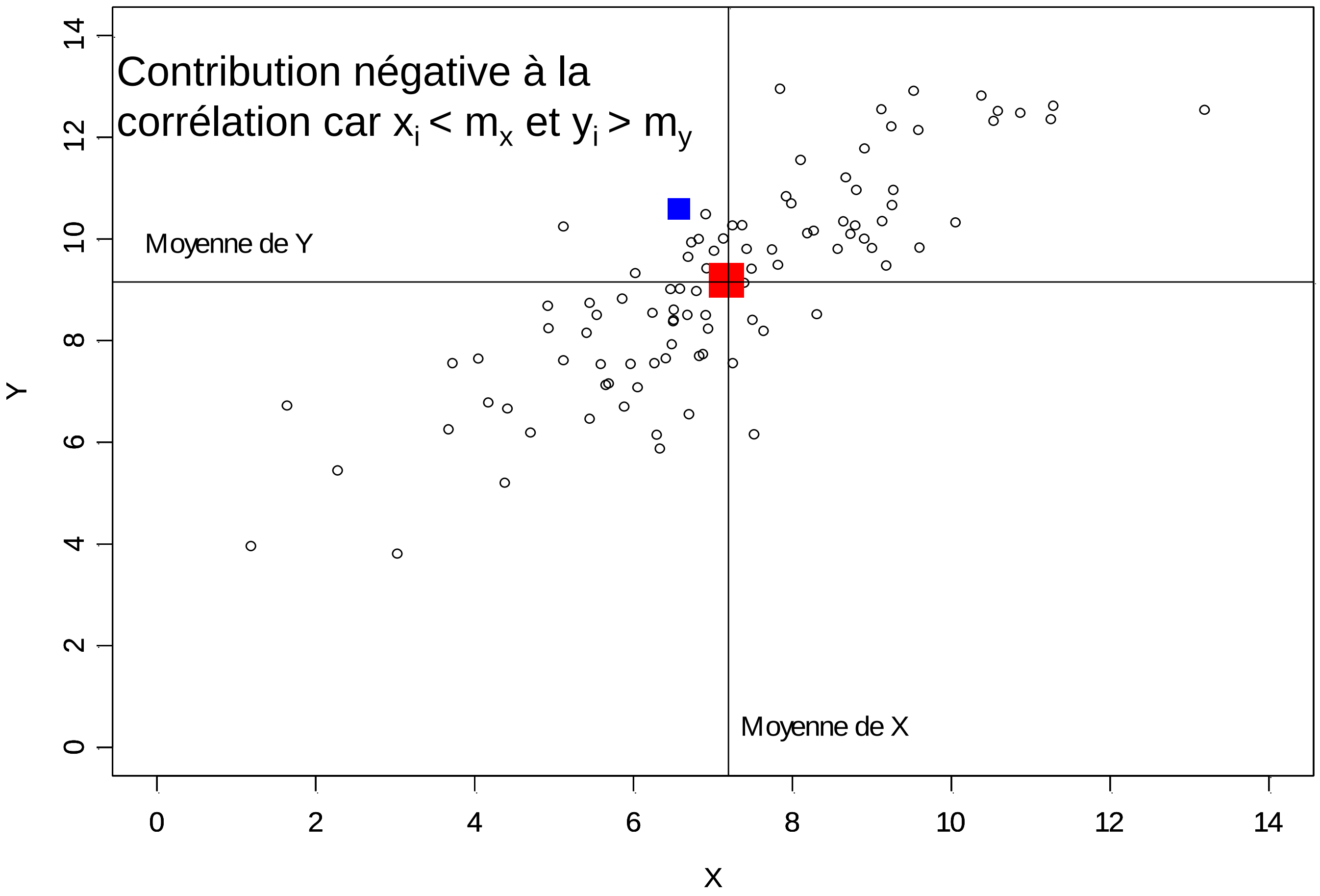


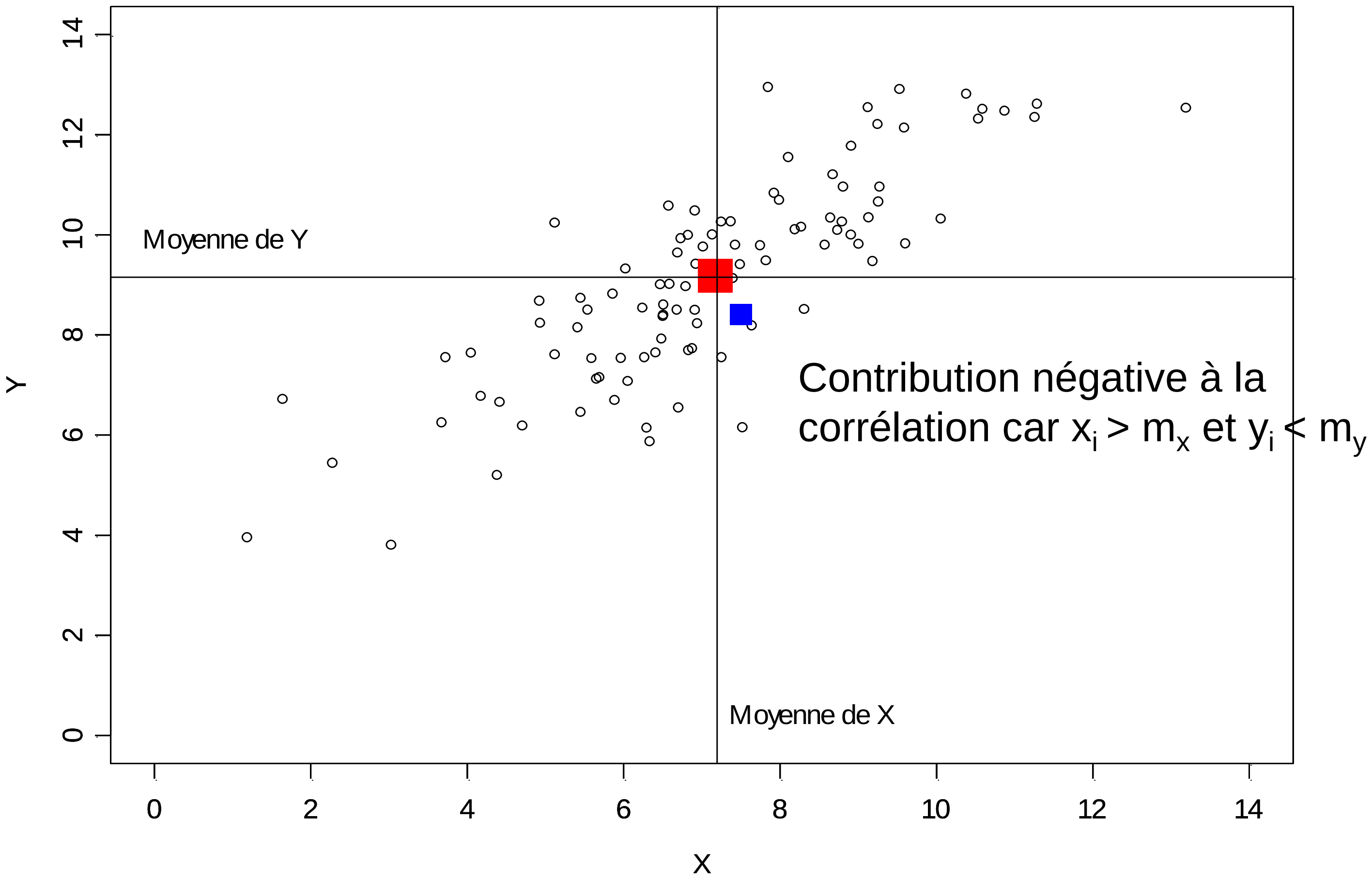



Moyenne de Y
Moyenne de X
Contribution positive à la
corrélation car $x_i > m_x$ et $y_i > m_y$
X
Y
0
2
4
6
8
10
12
14

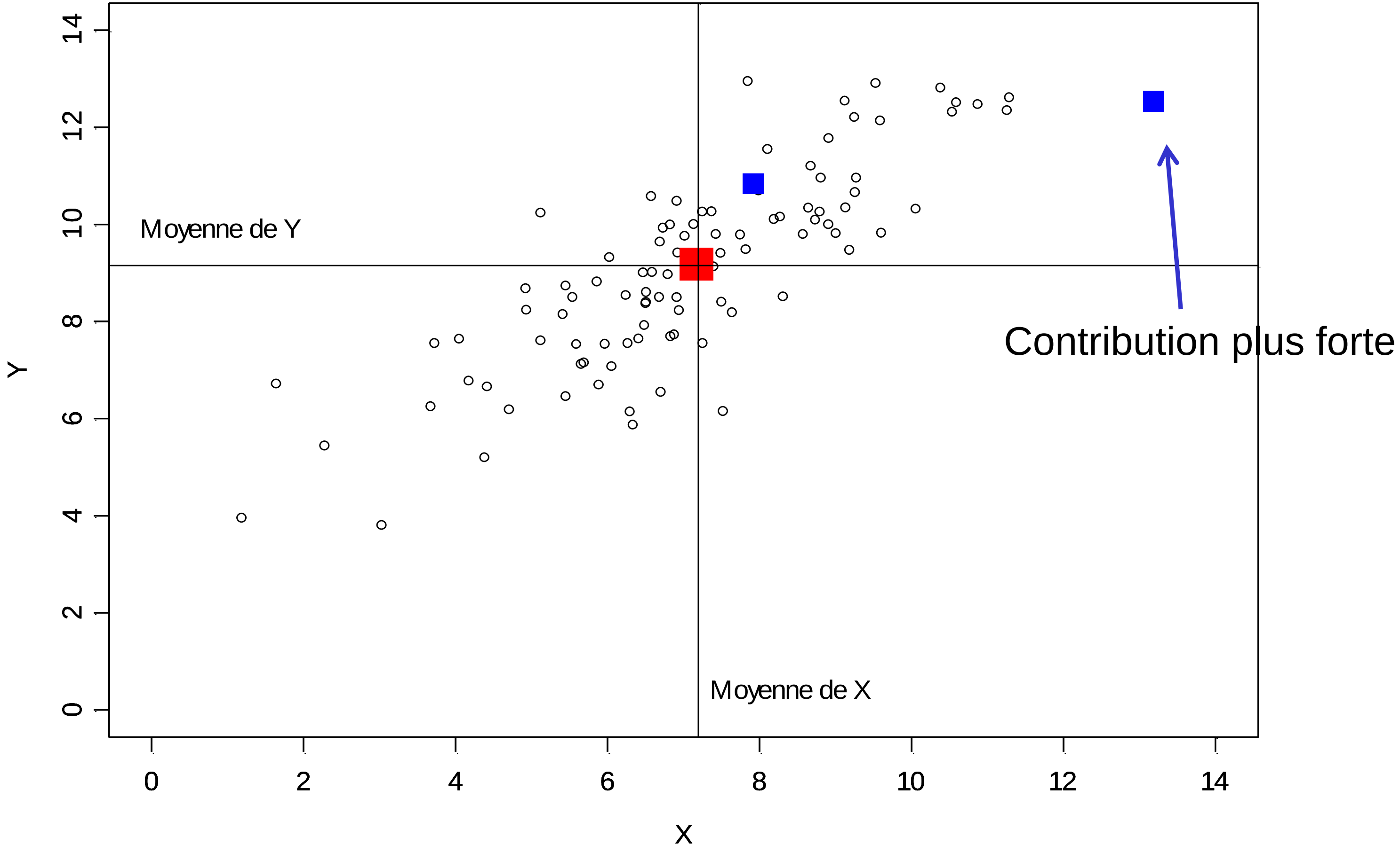


Contribution positive à la corrélation car $x_i < m_x$ et $y_i < m_y$

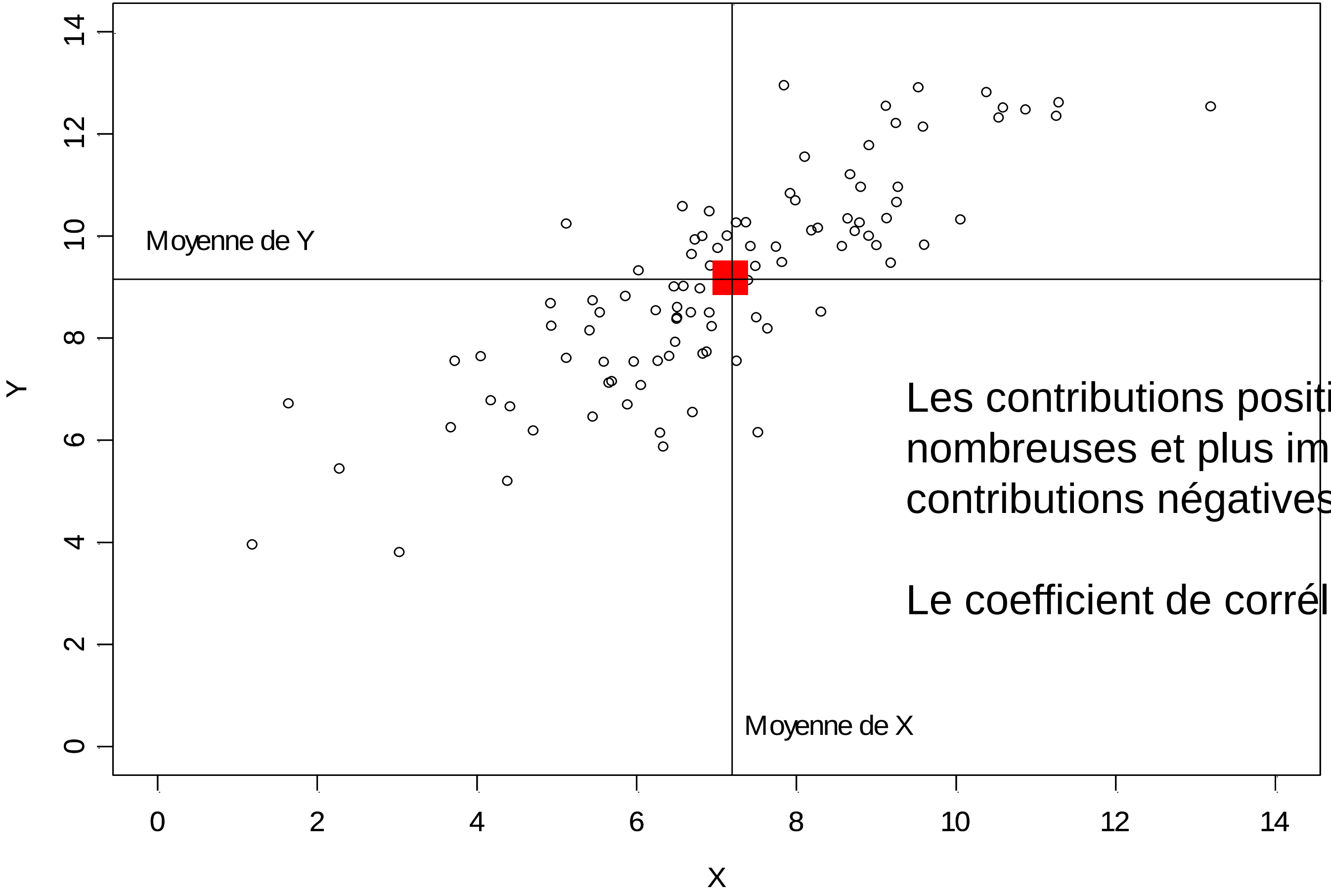

Contribution négative à la
corrélation car $x_i < m_x$ et $y_i > m_y$
Moyenne de Y
Moyenne de X
0
2
4
6
8
10
12
14
X
Y



Moyenne de Y
Moyenne de X
Contribution négative à la corrélation car $x_i > m_x$ et $y_i < m_y$
Y
X
0
2
4
6
8
10
12
14

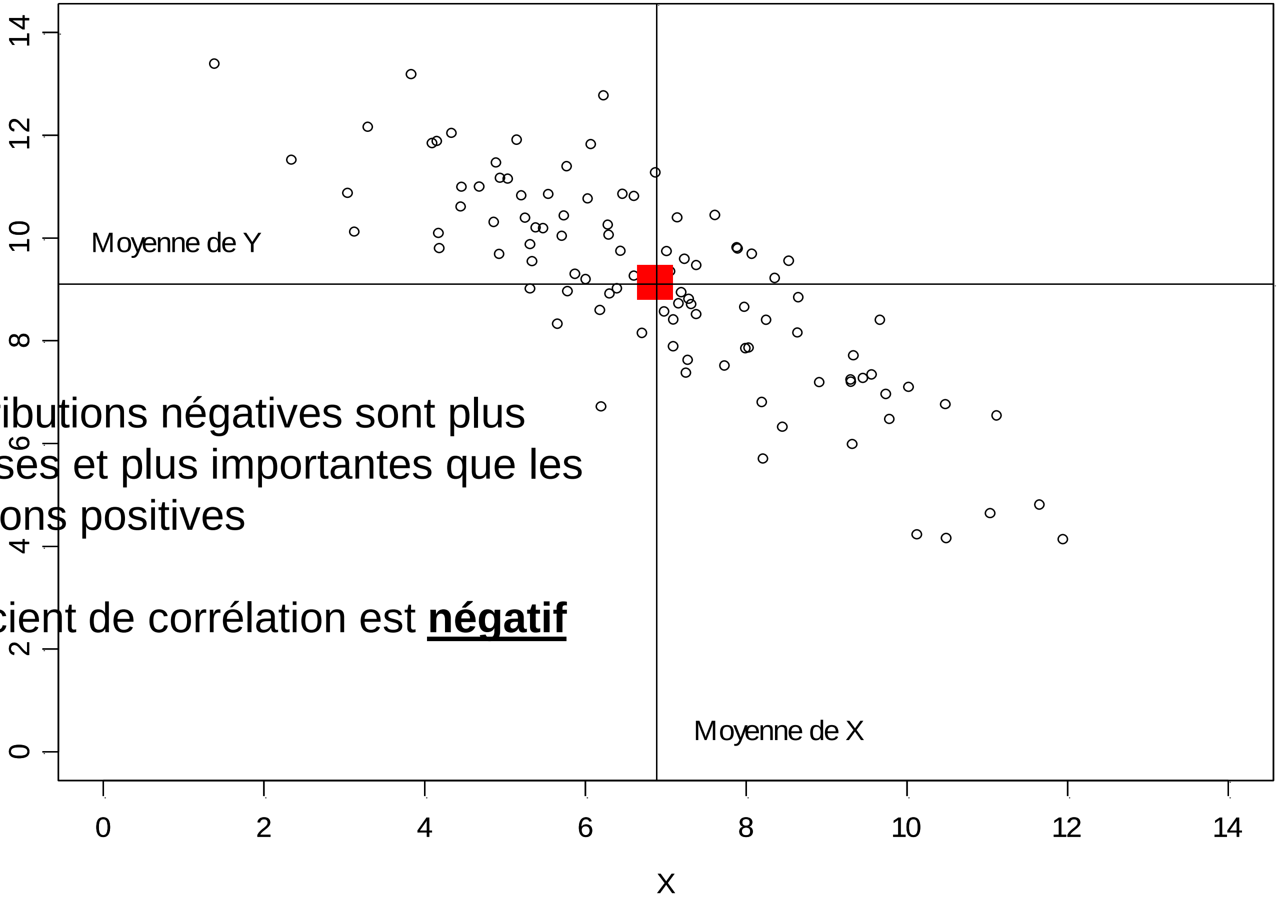


Moyenne de Y
Moyenne de X
Contribution plus forte
Y
X
0
2
4
6
8
10
12
14



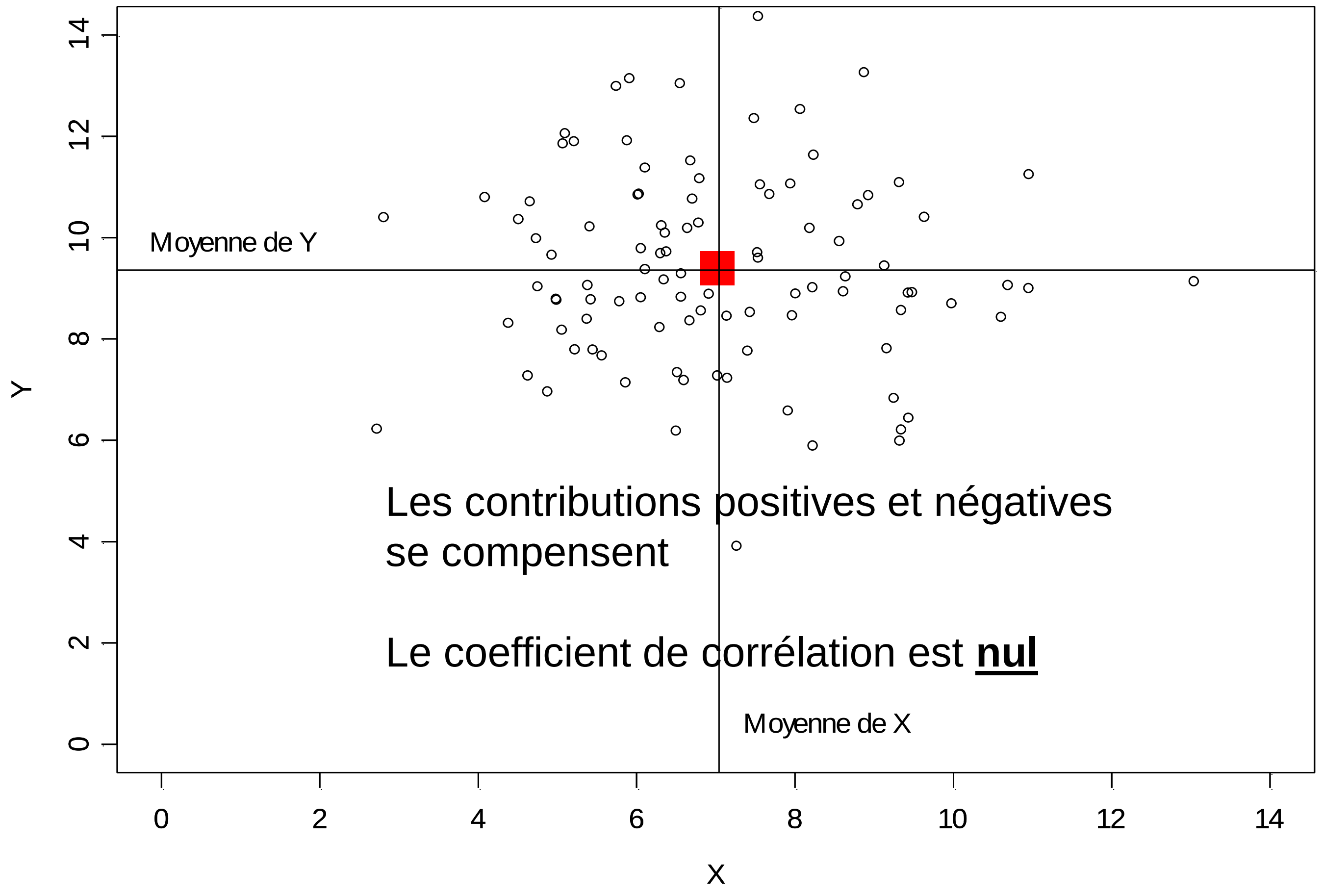
Les contributions positives sont plus nombreuses et plus importantes que les contributions négatives

Le coefficient de corrélation est **positif**



Les contributions négatives sont plus nombreuses et plus importantes que les contributions positives

Le coefficient de corrélation est **négatif**


Moyenne de Y
Moyenne de X
Les contributions positives et négatives
se compensent
Le coefficient de corrélation est **nul**
X
Y
0
2
4
6
8
10
12
14

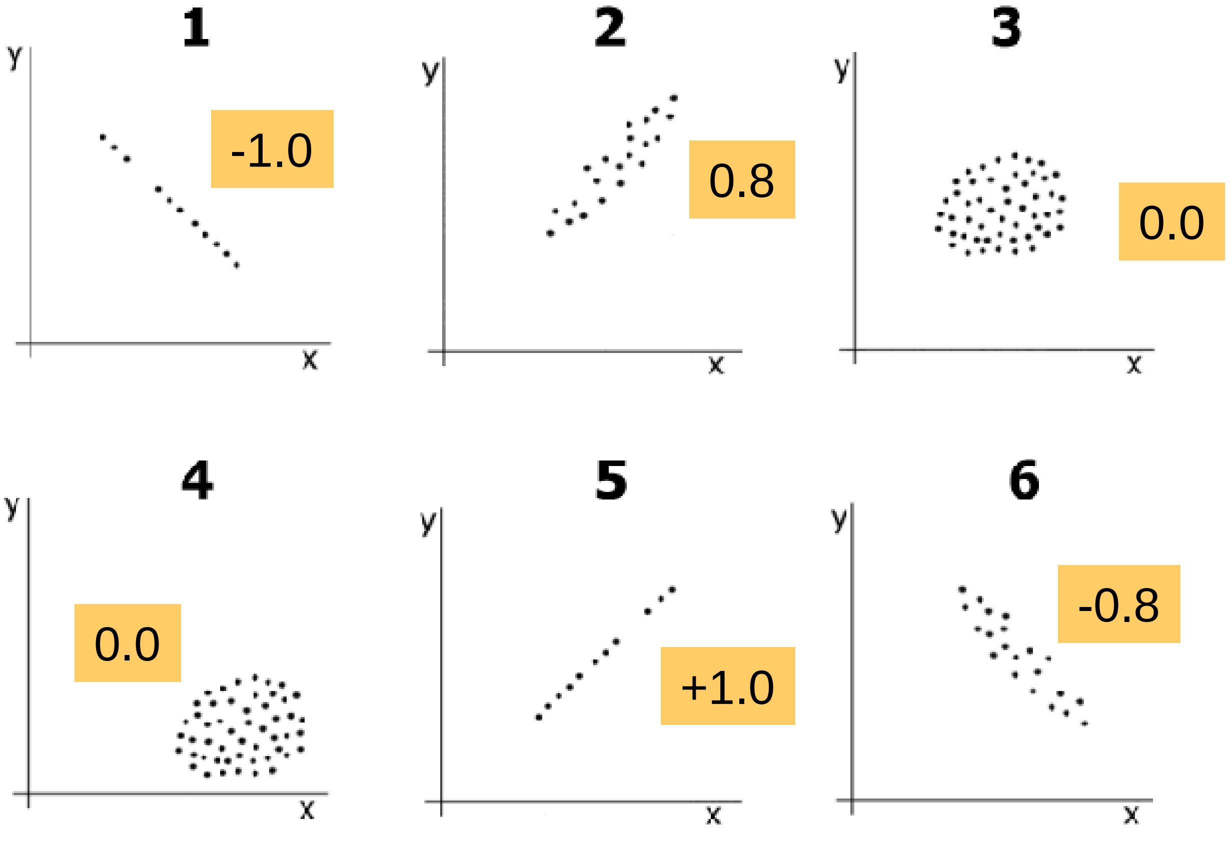
Coefficient de corrélation r

- Mesure le degré d'association **<u>linéaire</u>** entre 2 variables continues

- ***r*** varie entre –1 et +1:
  - –1.0 parfaite corrélation négative
  - –0.5 corrélation négative moyenne
  - 0.0 corrélation nulle
  - 0.5 corrélation positive moyenne
  - 1.0 parfaite corrélation positive

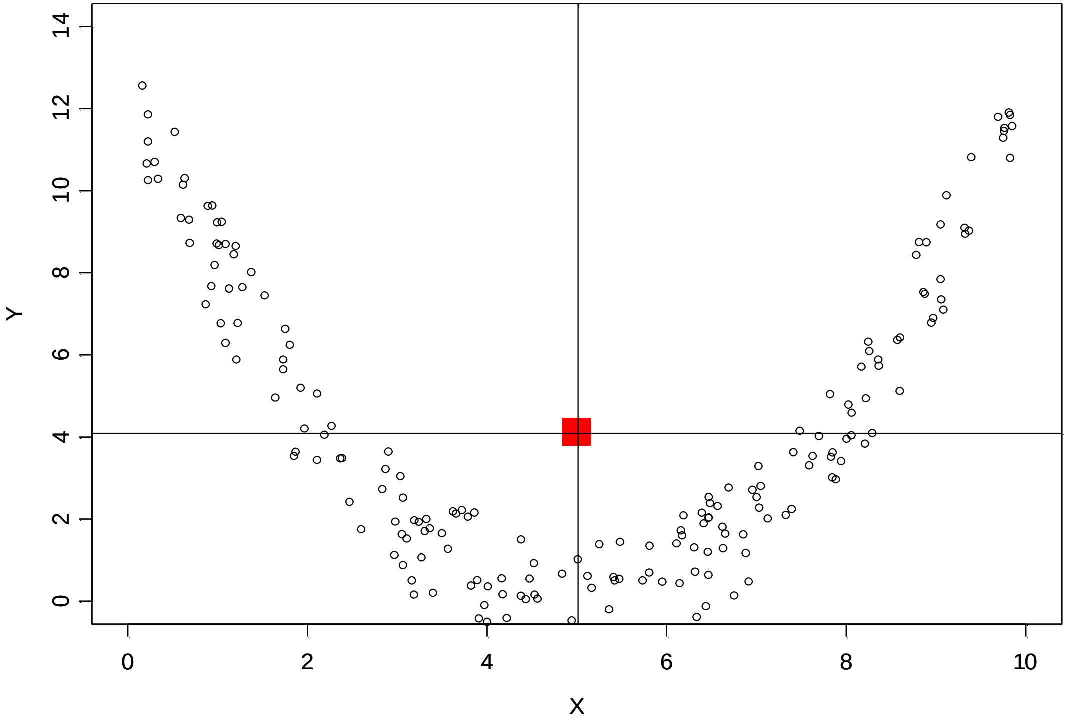
- Attention! si l'association existe, mais n'est pas linéaire,
  r n'est pas interprétable correctement

Combien vaut r ?

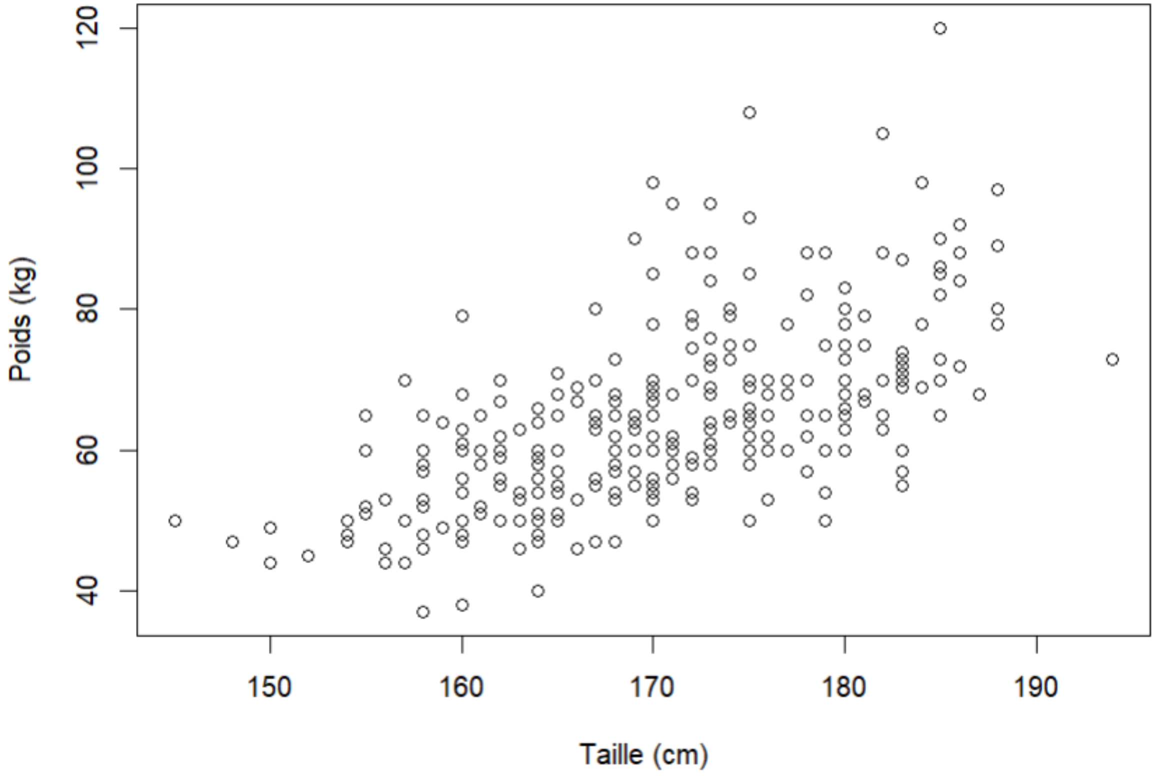


Association non linéaire

Les contributions négatives et positives se compensent: ***r***=0
Mais pourtant, il existe un lien entre X et Y



Retour à l'exemple Poids/Taille



r = 0.62
p < 0.001
r^2 = 0.39

r est un estimateur du paramètre ρ

r est affecté d'une erreur d'estimation, ou erreur-type, qui permet de calculer une valeur p

H_0: $\rho = 0$

Corrélation entre poids et taille

- Le poids et la taille sont corrélés parmi les étudiant-es en médecine de 1ère année

- La corrélation est positive, forte (r=0.62), et statistiquement significative ($p<0.001$)

Quel coefficient de corrélation?

- **Pearson**
  - Le plus fréquemment utilisé
  - Correspond à la formule de la diapo 10
  - Idéal lorsque les 2 variables ont une distribution normale

0.62

- **Spearman**
  - Utilise les rangs, plutôt que les valeurs exactes
  - A utiliser lorsque les distributions des variables sont asymétriques ou biscornues

0.66

Interprétation identique!

Modélisation

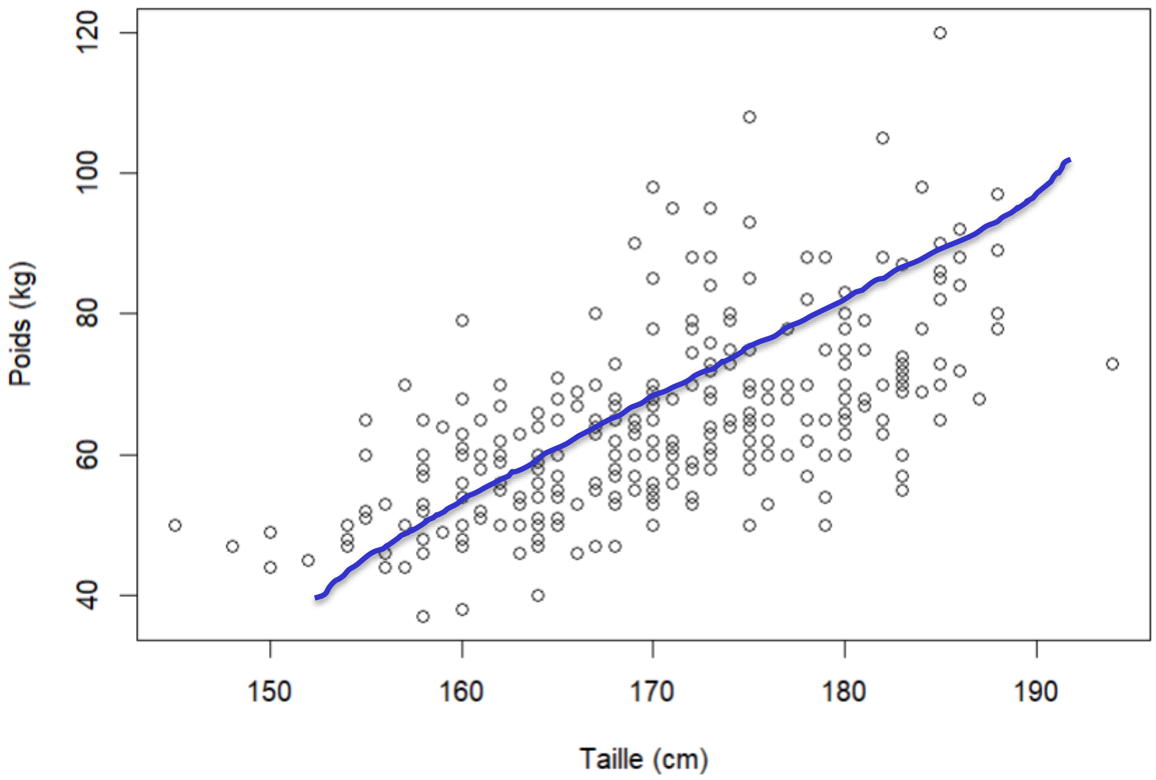
- Buts de la modélisation:
  - comprendre un phénomène complexe
  - décrire en résumant
  - prédire le comportement du système

- Tous les modèles sont faux! Mais certains sont utiles

- Utilités de la modélisation en recherche clinique:
  - détecter des associations
  - prédire chez des patients la valeur d'une variable difficile à mesurer

Quelle relation entre taille et poids ?

Nuage de points (scatter plot)



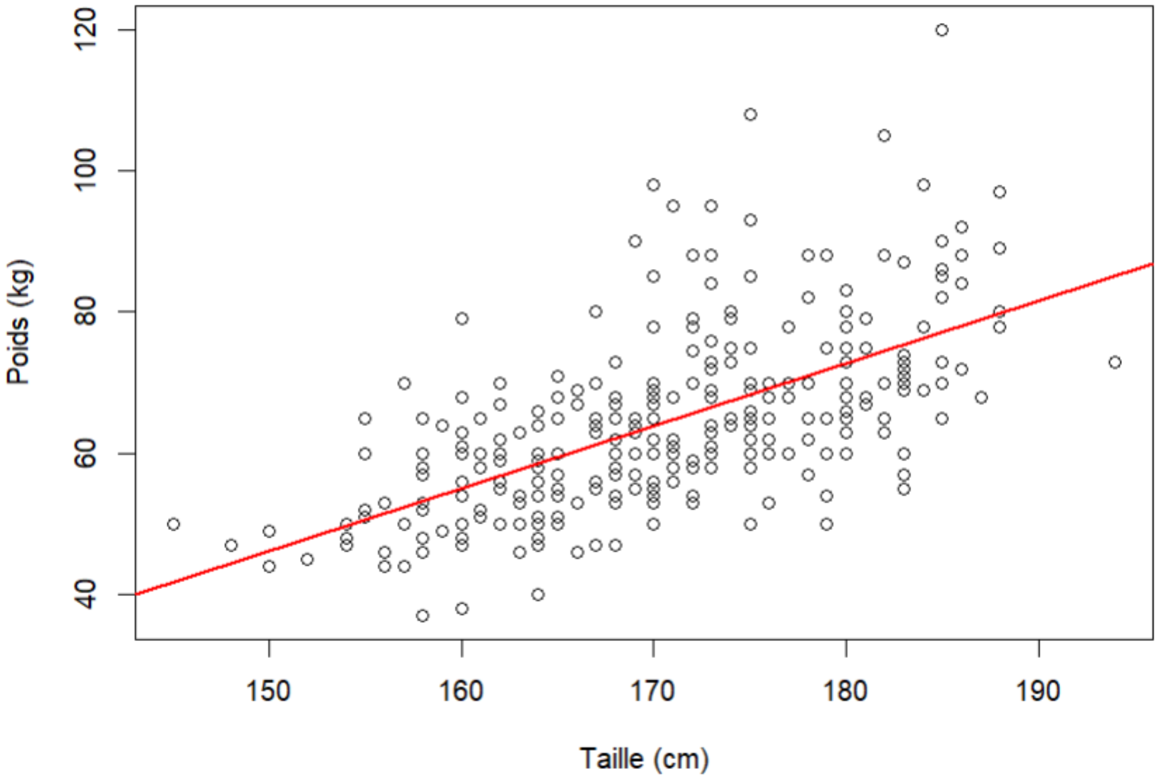
Le poids semble augmenter avec la taille et de manière linéaire

Une droite semble appropriée pour représenter cette relation

Mais quelle droite ?

La meilleure droite possible est

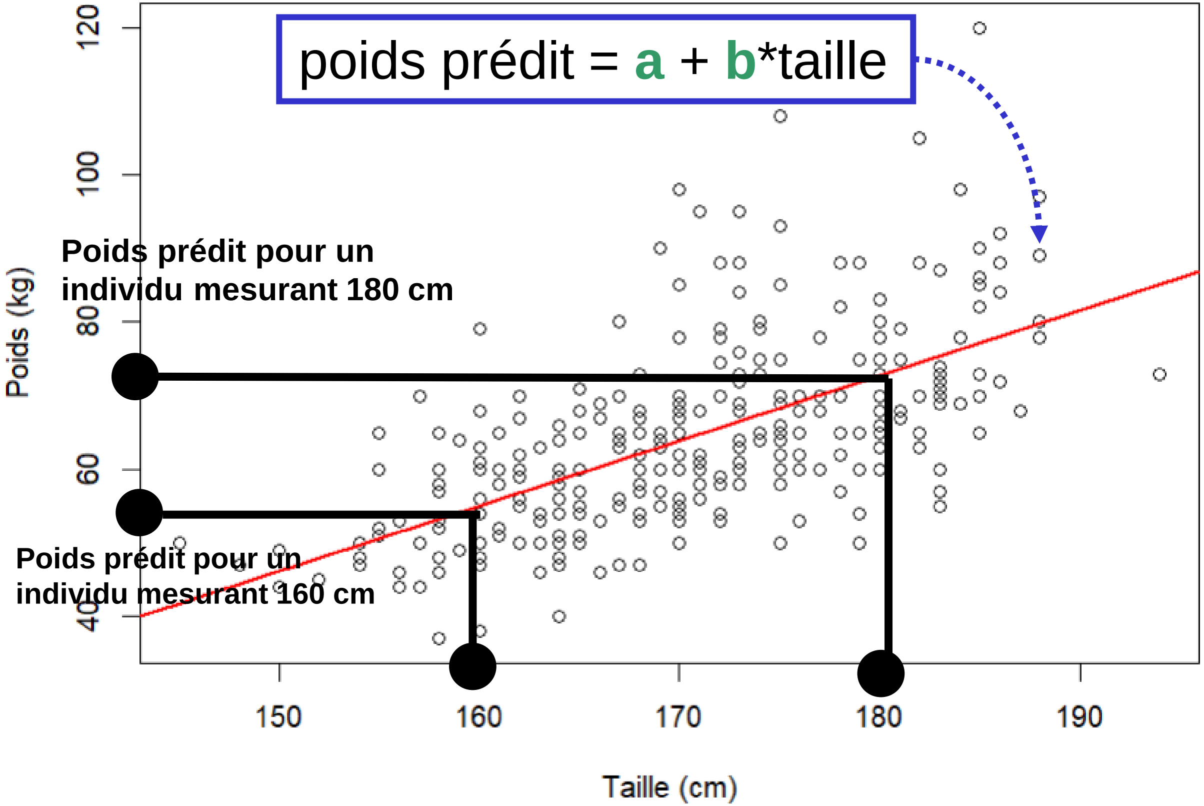
Nuage de points (scatter plot)



Elle est obtenue par **régression linéaire**

On l'appelle **droite de régression**

Interprétation de la droite de régression

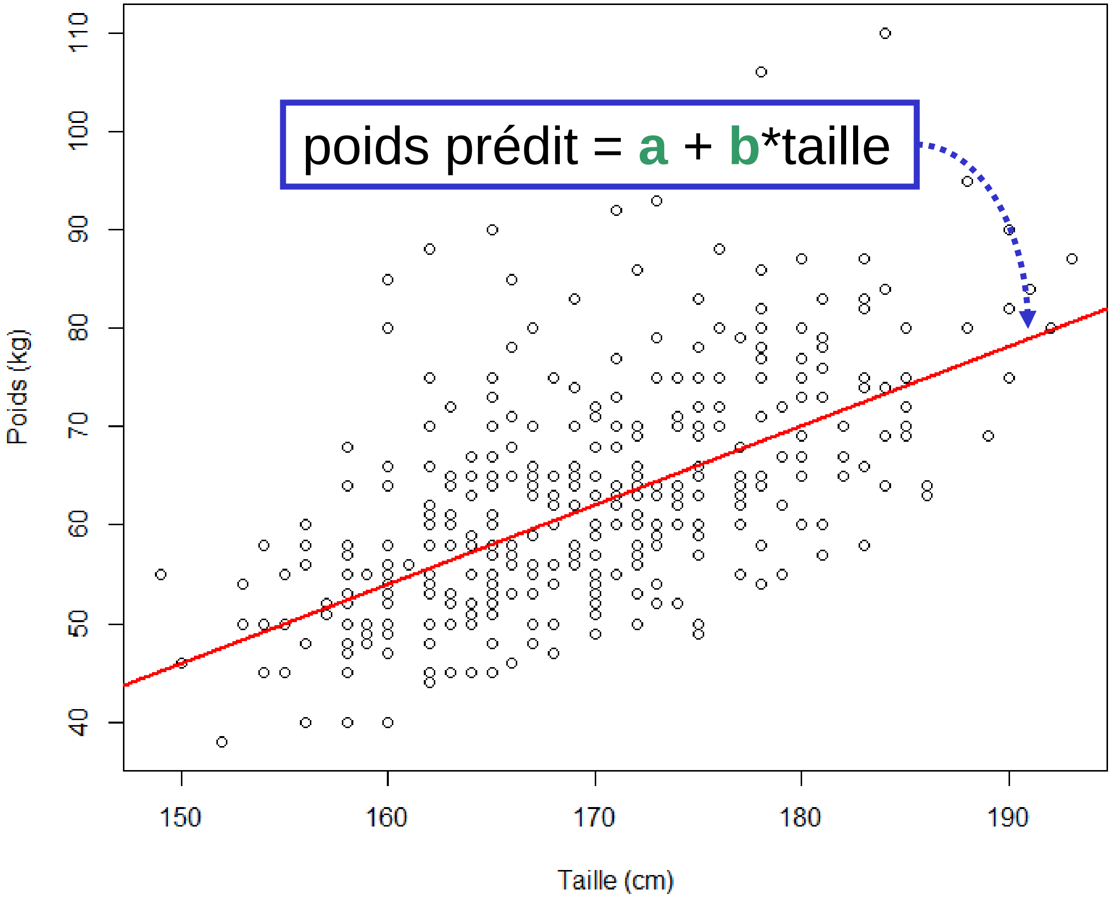


La valeur du poids **prédite** par le modèle pour une personne de 160 cm est (en kg):
a + 160 **b**

La valeur du poids **prédite** par le modèle pour une personne de 180 cm est (en kg):
a + 180 **b**

a : intercept (constante)
b : pente de la droite de régression

Interprétation de la droite de régression



a: poids prédit si taille = 0 cm

b: poids (kg) supplémentaire par cm de taille additionnel

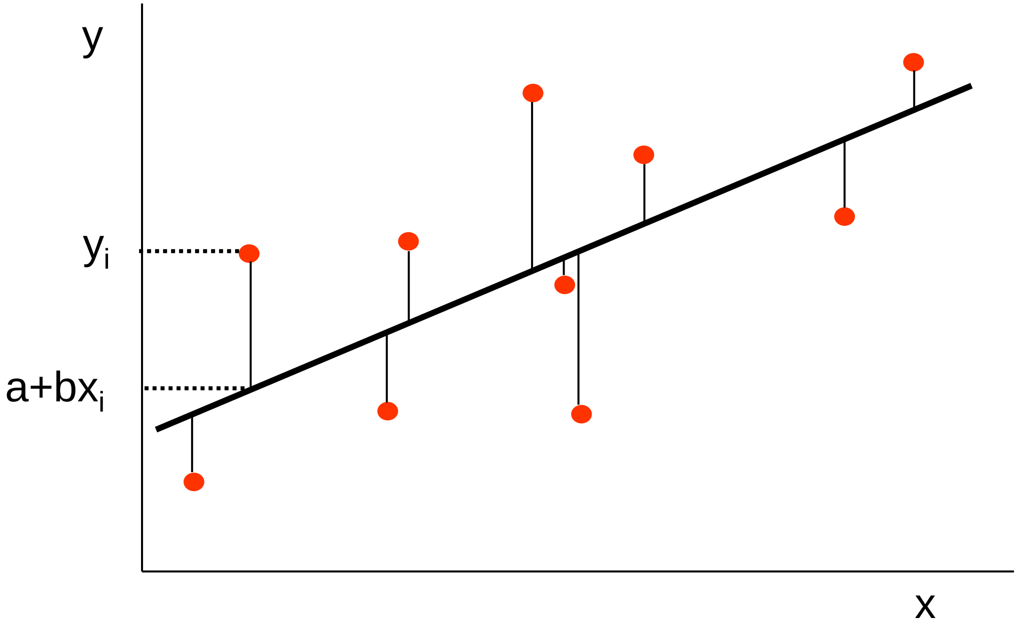
b: décrit l'association entre le poids et la taille

Absence d'association: b=0

La régression linéaire est la méthode qui permet d'estimer a et b à partir des données observées

Principe de la régression linéaire

- Droite qui minimise la somme des écarts mis au carré
(= qui minimise la variance résiduelle)



$y_i - (a+bx_i)$: différence entre la valeur observée (y_i) et la valeur prédite par le modèle ($a+bx_i$)

Ces différences sont appelées les résidus

Il faut trouver **a** et **b** qui minimisent la fonction:

$$\sum (y_i - (a+bx_i))^2$$

Poids/Taille: Résultat régression

Coefficients de
régression

		Estimate	Std. Error	t value	Pr(>|t|)
a	(Intercept)	-86.34215	10.88306	-7.934	4.26e-14
b	Taille	0.88429	0.06388	13.843	< 2e-16

Pente:
pour chaque cm de taille additionnel, le poids augmente en moyenne de 0.88 kg

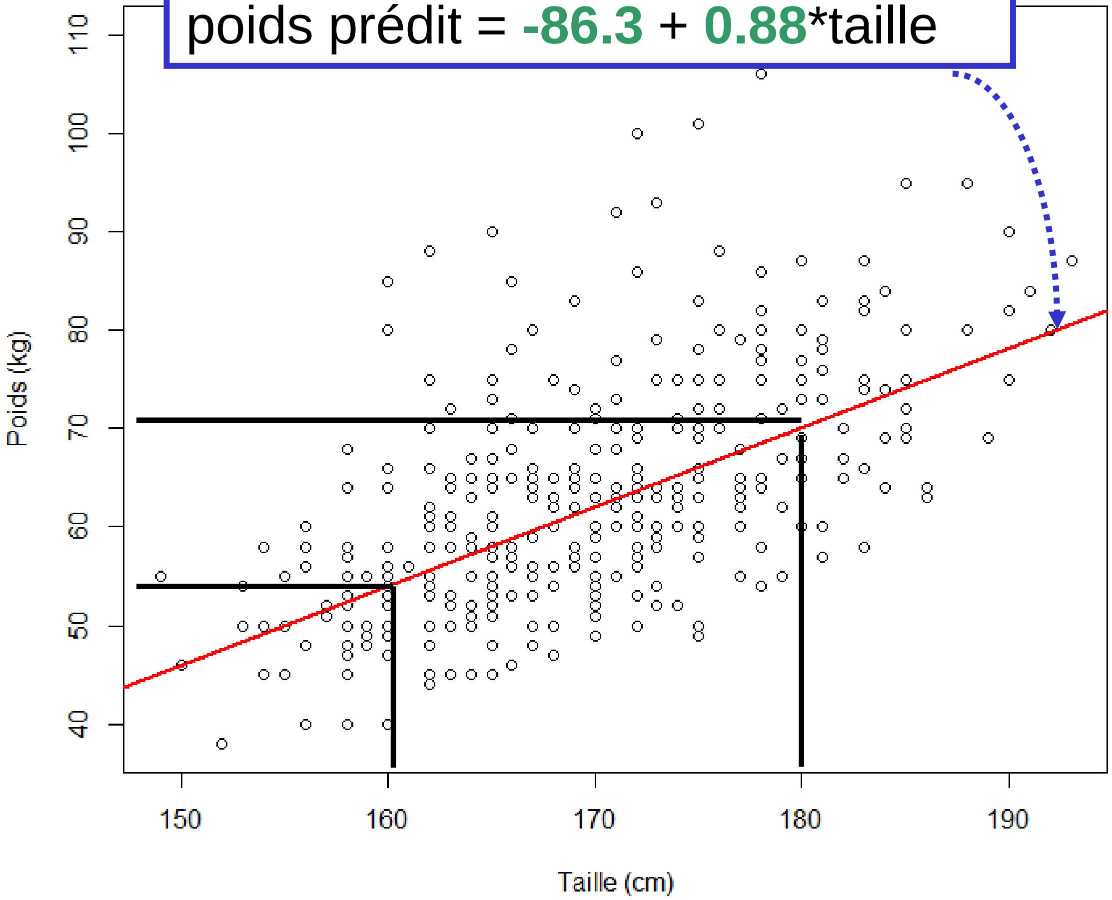
Constante:
poids prédit si taille = 0 cm

Equation pour prédire le poids en fonction de la taille

$$\boxed{\text{Poids prédit (kg)} = -86.3 + 0.88 \times \text{Taille (cm)}}$$

Poids/Taille: Résultat régression

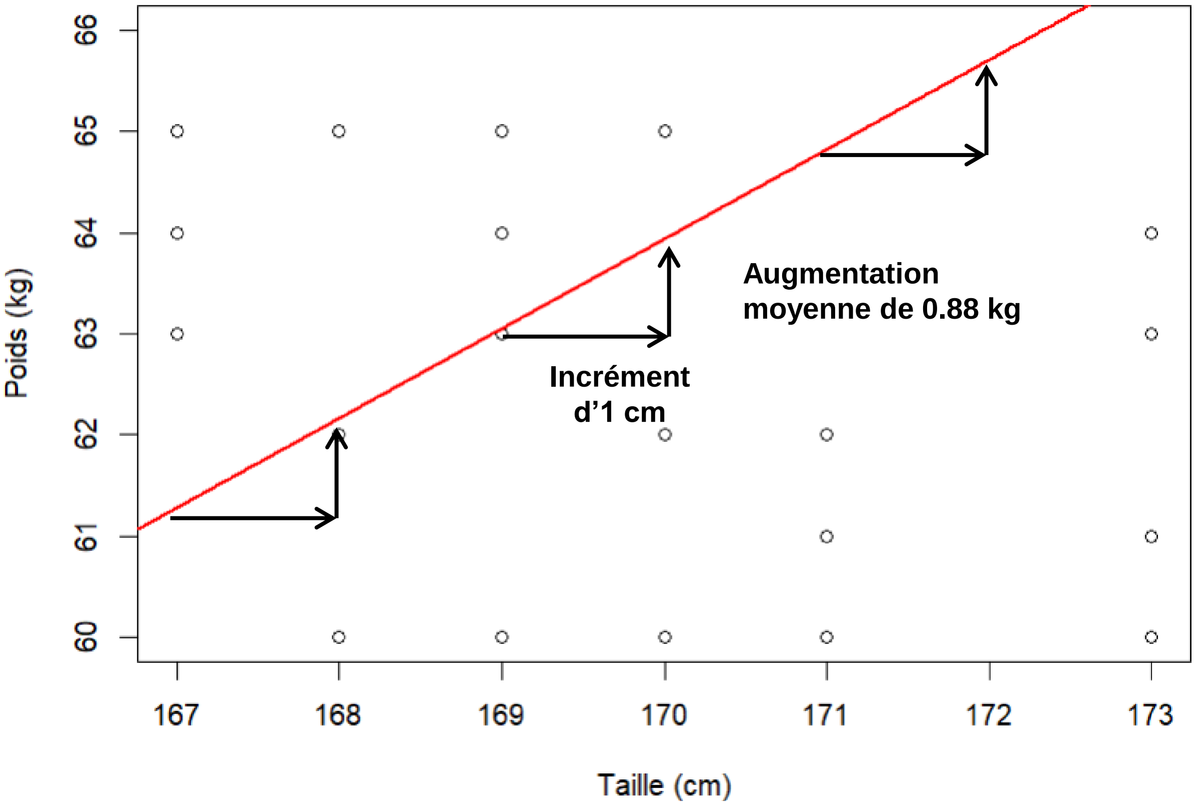
poids prédit = -86.3 + 0.88*taille



La valeur du poids **prédite** par le modèle pour une personne de 160 cm est (en kg):
-86.3 + 160 x 0.88 = 54.5

La valeur du poids **prédite** par le modèle pour une personne de 180 cm est (en kg):
-86.3 + 180 x 0.88 = 72.1

Poids/Taille: Résultat régression

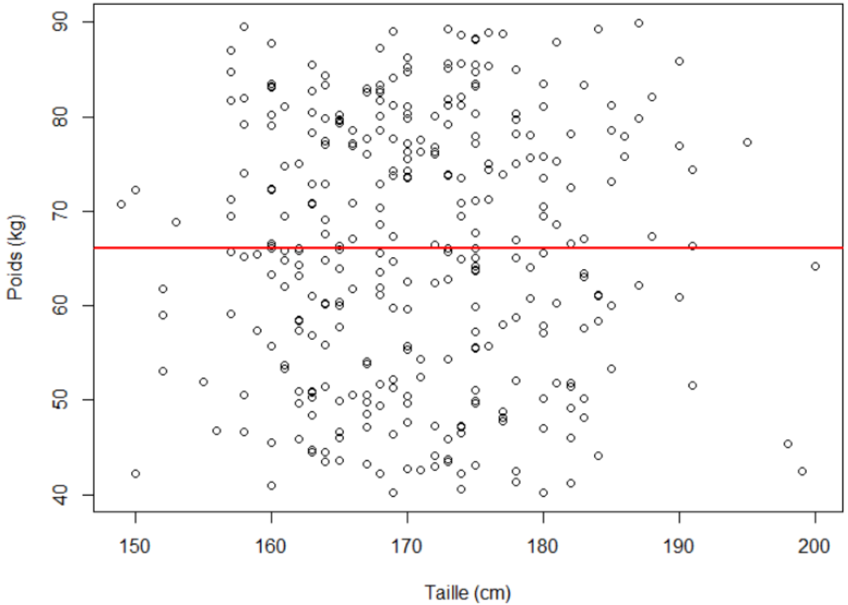


Pente: 0.88 kg/cm

Pour chaque centimètre additionnel de taille, le poids augmente en moyenne de 0.88 kg

Absence d'association

- Si le poids n'est pas associé à la taille dans la population, on s'attend à observer dans l'échantillon:
  - poids moyen identique quelle que soit la taille
  - pente égale à 0 (droite de régression horizontale)
  - changement de poids moyen par centimètre de taille supplémentaire = 0 kg/cm



$$\boxed{\text{Poids prédit} = a + 0 \cdot \text{taille}}$$

Poids/Taille: Résultat régression

Erreurs d'estimation
(erreurs types)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -86.34215   10.88306  -7.934 4.26e-14
Taille        0.88429    0.06388  13.843  < 2e-16
```

Intervalle de confiance à 95% de la constante -86.3:

Intercept +/- 2 Erreur type = -86.3 +/- 2*10.9 = -108.1 à -64.5

Intervalle de confiance à 95% de la pente 0.88:

Pente +/- 2 Erreur type = 0.88 +/- 2*0.06 = 0.76 à 1.00

Une pente nulle est peu
compatible avec les données

Poids/Taille: Résultat régression

Tests statistiques

Statistiques de test

Valeurs p

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-86.34215	10.88306	-7.934	4.26e-14
Taille	0.88429	0.06388	13.843	< 2e-16

Hypothèses nulles testées:

H_0: constante = 0
H_0: pente = 0 (absence d'association entre le poids et la taille)

Association poids-taille

- La taille permet de prédire le poids selon l'équation:

$$\boxed{\text{Poids prédit} = -\ 86.3 + 0.88 * \text{taille}}$$

- L'augmentation de poids par centimètre supplémentaire de taille est 0.88 kg/cm
- L'augmentation de poids par centimètre supplémentaire a un intervalle de confiance à 95% de 0.76 à 1.00 kg/cm
- L'association entre taille et poids est statistiquement significative ($p<0.001$)

Unités des coefficients de régression

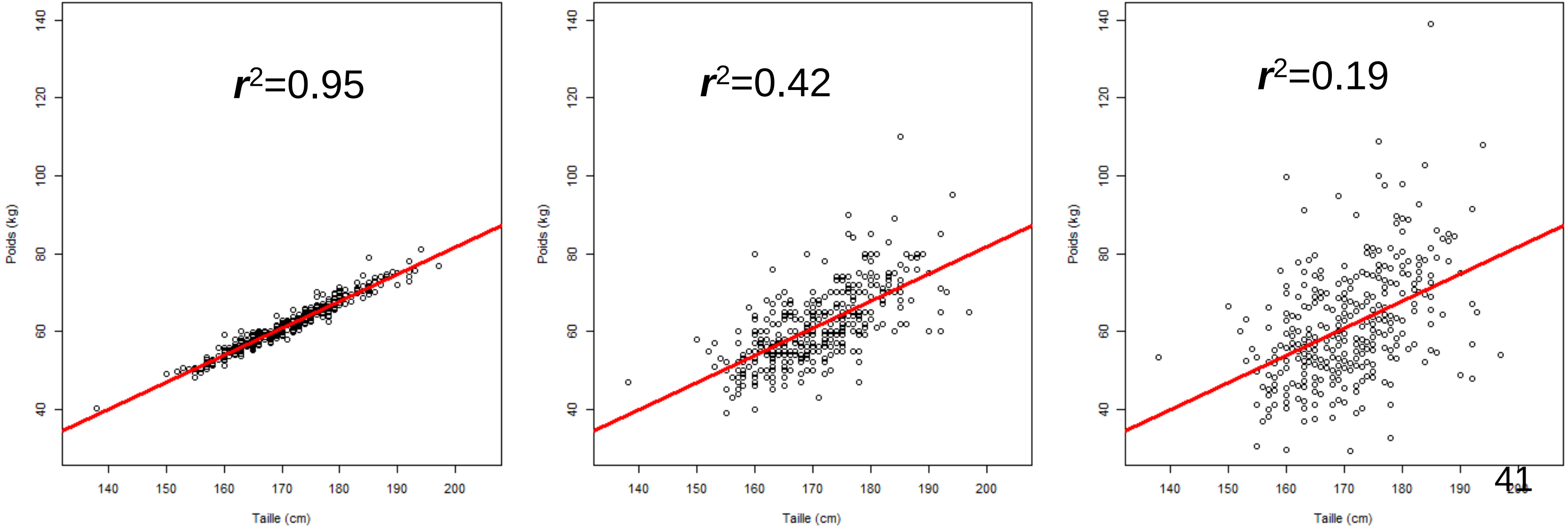
- Les coefficients a et b dépendent des unités des variables étudiées
- Exemple Taille (cm) / Poids (kg)
  - a = -86.3 kg
  - b = 0.88 kg/cm
- Si la taille est exprimée en <u>mètre</u> et le poids en kilogramme
  - a = -86.3 kg
  - b = 88 kg/m
- Si la taille est exprimée en centimètre et le poids en <u>gramme</u>
  - a = -86 300 g
  - b = 880 g/cm

Comment expliquer la relation entre taille et poids ?

- Grande taille cause poids plus élevé – **causalité**
- Poids plus élevé cause grande taille – **causalité inverse**
- Une tierce variable explique la relation – effet de **confusion**
- Il n’y a pas de relation en réalité – ce qu’on voit est dû à un **biais** méthodologique
- Il n’y a pas de relation en réalité – ce qu’on voit est dû au **hasard**

Coefficient de détermination R^2

- r^2 est le carré du coefficient de corrélation de Pearson
- Mesure la qualité de la prédiction par un modèle de régression linéaire
- r^2 est entre 0 (X n'explique rien de Y) et 1 (prédiction parfaite)
- Plus les résidus sont faibles (= plus les valeurs observées sont proches des valeurs prédites), plus r^2 est proche de 1



Coefficient de détermination R^2

r^2 est la proportion de la variance commune aux 2 variables

Exemple: poids et taille, r^2 = 0.39

- le poids explique un peu moins de la moitié (39%) de la variance de la taille
- la taille explique environ un peu moins de la moitié (39%) de la variance du poids

Attention: r^2 ne dit rien sur la causalité

Exemple

Kim *et al. BMC Nephrology* (2015) 16:131
DOI 10.1186/s12882-015-0131-4

BMC Nephrology

RESEARCH ARTICLE **Open Access**

CrossMark

Cross-sectional association of volume, blood pressures, and aortic stiffness with left ventricular mass in incident hemodialysis patients: the Predictors of Arrhythmic and Cardiovascular Risk in End-Stage Renal Disease (PACE) study

Esther D. Kim[1], Stephen M. Sozio[2,3], Michelle M. Estrella[2,3], Bernard G. Jaar[2,3,4], Tariq Shafi[2,3], Lucy A. Meoni[2,3,5], Wen Hong Linda Kao[2,3,6], Joao A. C. Lima[2] and Rulan S. Parekh[1,2,6,7,8*]

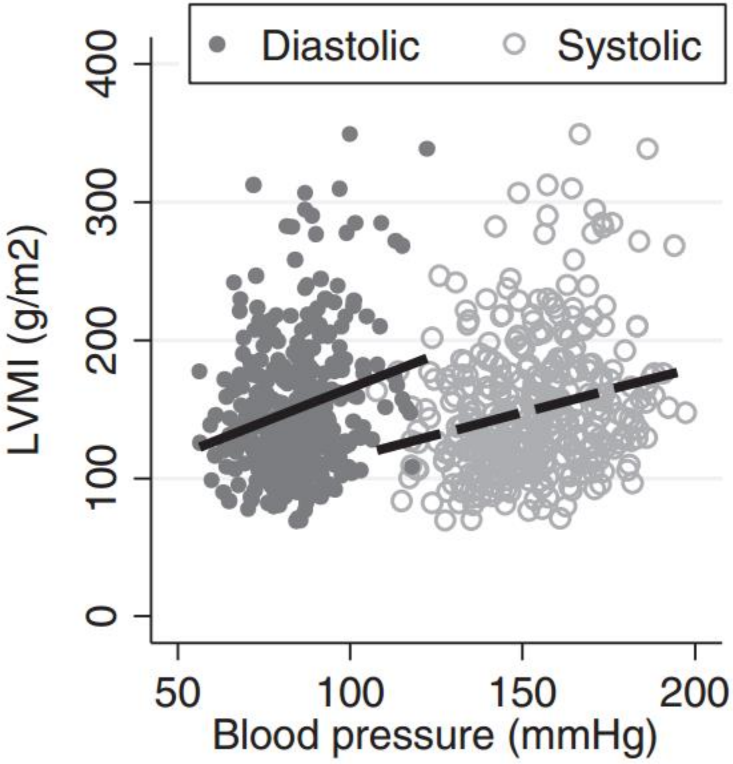
Exemple

Background: Higher left ventricular mass (LV) strongly predicts cardiovascular mortality in hemodialysis patients. Although several parameters of preload and afterload have been associated with higher LV mass, whether these parameters independently predict LV mass, remains unclear.

Methods: This study examined a cohort of 391 adults with incident hemodialysis enrolled in the Predictors of Arrhythmic and Cardiovascular Risk in End Stage Renal Disease (PACE) study. The main exposures were systolic and diastolic blood pressure (BP), pulse pressure, arterial stiffness by pulse wave velocity (PWV), volume status estimated by pulmonary pressures using echocardiogram and intradialytic weight gain. The primary outcome was baseline left ventricular mass index (LVMI).

Conclusions: Among a younger and incident hemodialysis population, higher systolic, diastolic, or pulse pressure, regardless of timing with dialysis, is most associated with higher LV mass.

Exemple



- ♦ Pente des droites de régression

"every 10 mmHg increase in systolic or diastolic BP was significantly associated with higher LVMI."

Tension artérielle:
- diastolique: 10.1, 95 % IC: 5.1 to 15.0
- systolique: 7.3, 95 % IC: 4.3 to 10.2

- ♦ Le coefficient de détermination r^2 est visiblement faible

Régression linéaire: résumé

- L’équation de régression linéaire **$Y = a + bX$** est un exemple simple de **modèle statistique**
  - relation entre Y et X supposée linéaire
  - Y modélisé avec X seulement
- Ce modèle examine une association linéaire entre 2 variables continues:
  - variable indépendante (X)
  - variable dépendante (Y)
- Questions de recherche:
  - est-ce que X est associé à Y ?
  - est-ce que X permet de prédire Y ?
- Principe: Établir une droite prédisant Y en fonction de X

Régression linéaire: résumé

- Conduit à l'estimation d'une **pente** qui capte l'association entre X et Y:
  - incrément de Y pour un incrément d'une unité pour X
  - pente positive: Y augmente quand X augmente
  - pente négative: Y diminue quand X augmente
  - pente nulle: Y ne varie pas avec X (**$=H_0$**)

- La valeur de la pente dépend des unités de Y et de X (ex: 0.81 kg/cm)

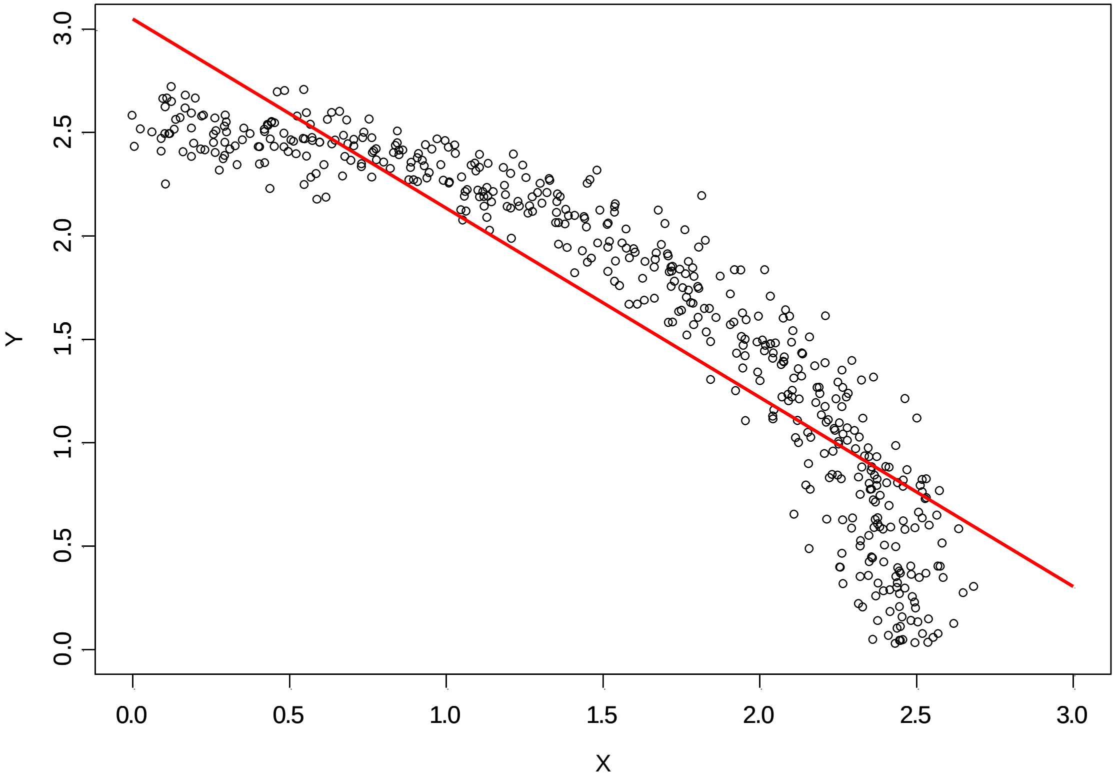
- r^2 est utile pour évaluer la capacité du modèle à prédire Y

- Attention: les modèles **$Y = a + bX$** et **$X = a + bY$** sont différents

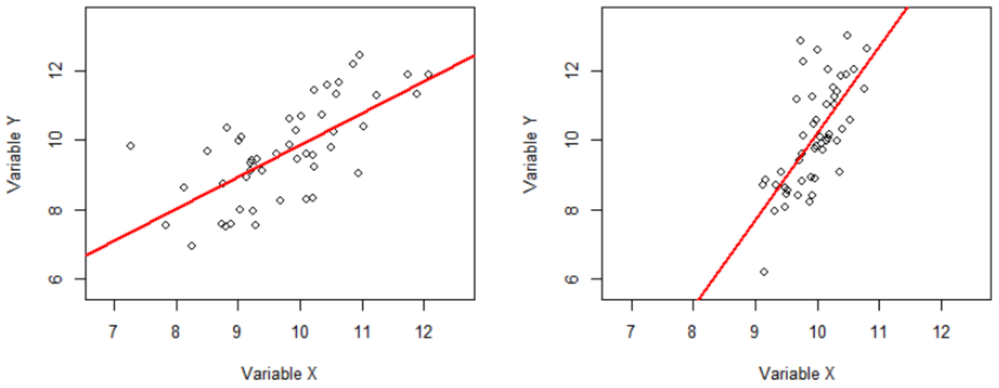
En principe…

- Pour qu'un modèle de régression linéaire soit valide, il faut que:
  - l'association entre X et Y doit être linéaire
  - les écarts $y_i - (a+bx_i)$ doivent avoir une distribution normale
  - la variance de Y doit être stable, quelle que soit la valeur de X

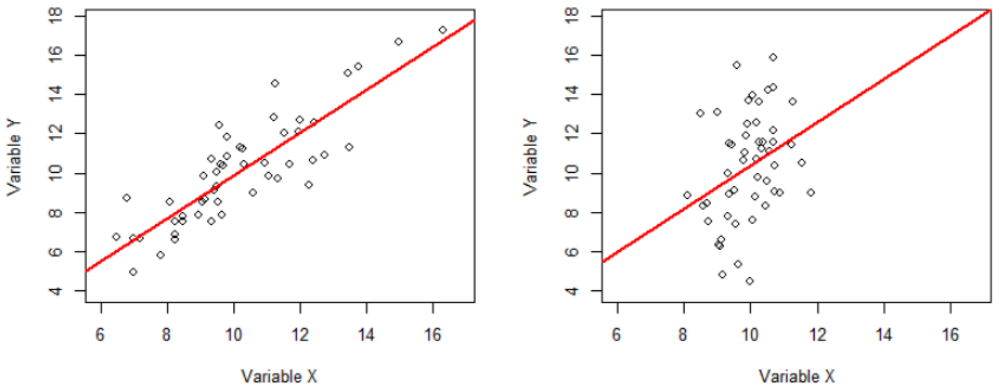
Exemple où la régression linéaire n'est pas adaptée



Corrélation versus régression



r identique mais pentes ≠



Pente identique mais ***r*** ≠

Corrélation versus régression

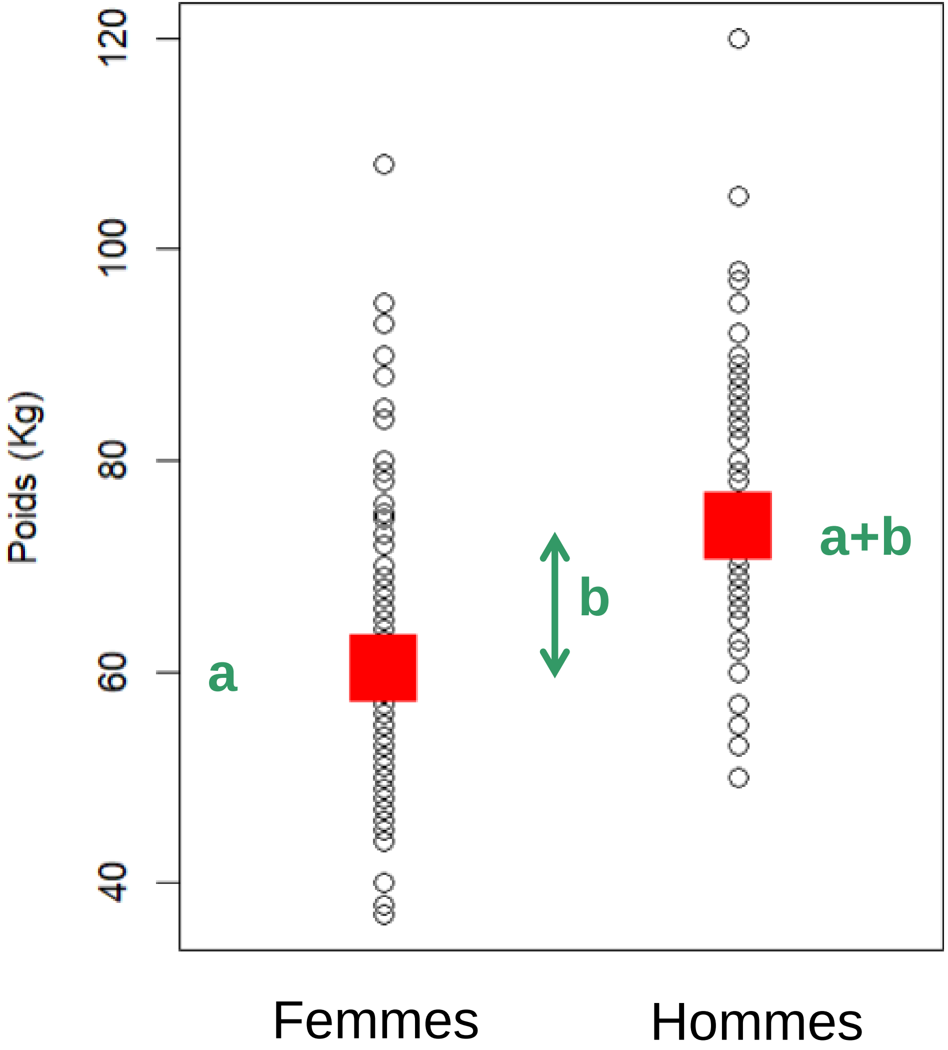
- **Corrélation** répond à la question:
  - décrit la force d'une association
  - traite les 2 variables de façon symétrique (cf. formule)
  - corrélation entre X et Y = corrélation entre Y et X
  - le coefficient de corrélation <u>ne dépend pas</u> des unités de mesure (il n'a pas d'unité)

- **Régression** répond à la question:
  - décrit l'influence d'une variable sur l'autre
  - distingue variable dépendante (Y) et variable indépendante (X)
  - coefficient de régression de Y sur X ≠ Coefficient de régression de X sur Y
  - les coefficients de régression <u>dépendent</u> des unités de mesure

Variable indépendante qualitative à 2 catégories

- L'équation de régression linéaire **$Y = a + bX$**
- On a vu le cas d'une variable indépendante quantitative continue (X: taille)
- Une variable indépendante peut aussi être qualitative
- Exemple: **Y = a + b Genre**
  - genre: 0 si Femme et 1 si Homme (par exemple)
  - Y prédit chez les femmes: $a + b \cdot 0 = a$
  - Y prédit chez les hommes: $a + b \cdot 1 = a + b$
  - a est le poids moyen chez les femmes
  - a+b est le poids moyen chez les hommes
  - b est la différence entre les poids moyens (hommes – femmes)

Poids moyen
chez les
femmes
(60.4 kg)



Poids moyen
chez les
hommes
(73.9 Kg)

Poids/genre: résultat régression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	60.3584	0.7486	80.63	<2e-16
GenreHomme	13.4904	1.4097	9.57	<2e-16

- Poids prédit par le modèle:
  - pour une femme: 60.4 kg
  - pour un homme : 60.4 + 13.5 = 73.9 kg
- Différence de poids moyen entre un homme et une femme: **b**=13.5 kg
- L'hypothèse nulle **b**=0 (poids moyen identique pour une femme et un homme) est rejetée ($p<0.001$)
  - hypothèse nulle identique à celle d'un test de Student
- r^2 = 0.23: le genre explique 23% de la variance du poids

Régression linéaire multiple

- Modélisation du poids en fonction de la taille et du genre
  - variable dépendante: Poids (kg)
  - variables indépendantes: Taille (cm), Genre (0:femme, 1:homme)

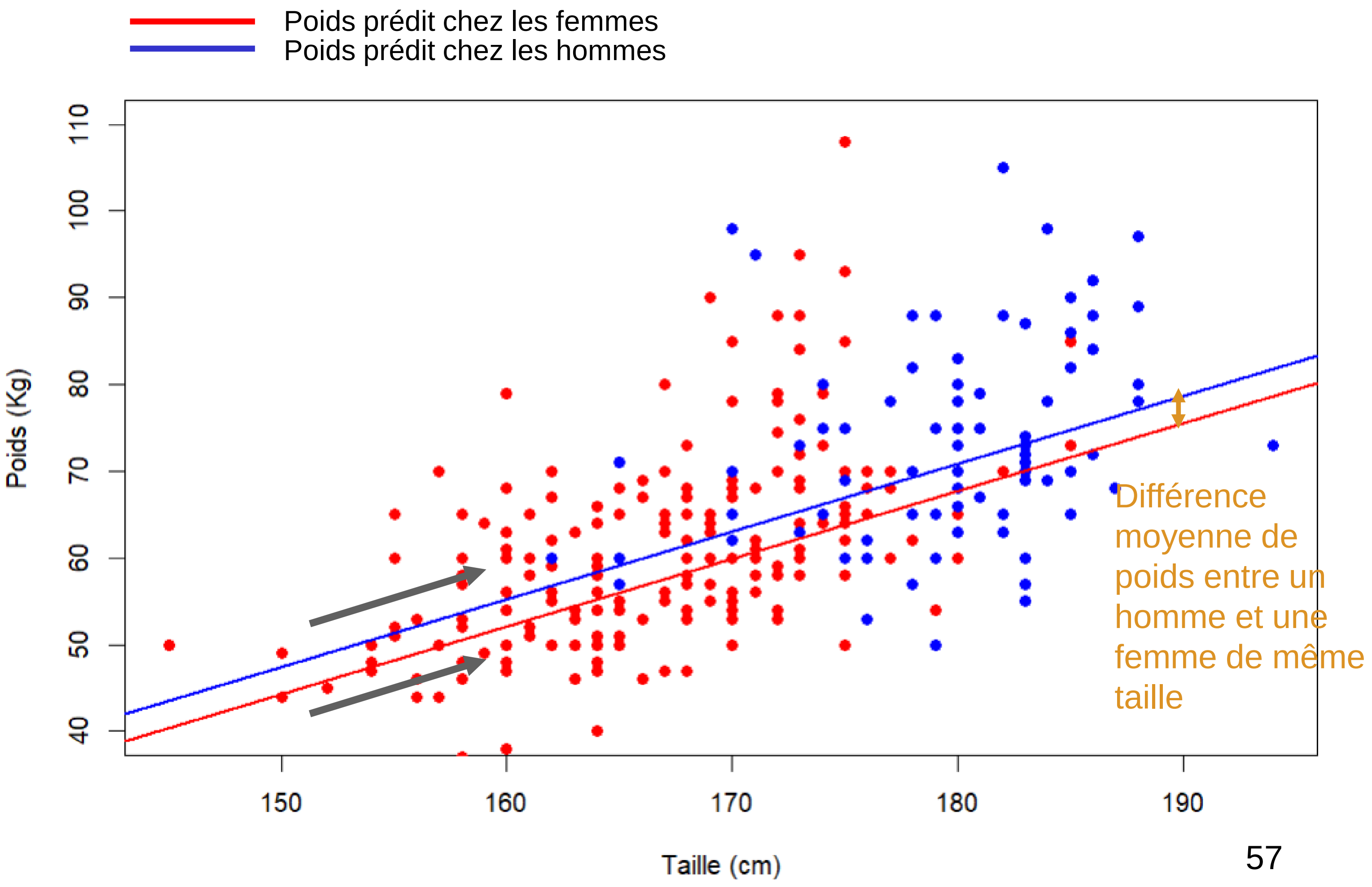
- Equation de régression linéaire **Poids = a + b Taille + c Genre**
- Poids prédit:
  - pour une femme de 170 cm = **a** + 170 **b**
  - pour un homme de 170 cm = **a** + 170 **b** + **c = (a+c)** + 170 **b**

- **a** : poids prédit chez une femme mesurant 0 cm
- **b** : changement du poids prédit par cm de taille supplémentaire chez des individus de **même genre**
- **c** : différence de poids prédits entre un homme et une femme de **même taille**

Régression multiple: exemple Poids

		Estimate	Std. Error	t value	Pr(>|t|)
a	(Intercept)	-72.8947	9.4134	-7.744	4.16e-14
b	Taille	0.7812	0.0608	12.849	< 2e-16
c	Genre	3.2040	1.1935	2.685	0.00746

- Chez des individus de même genre, le poids augmente en moyenne de 0.78 kg (IC95% 0.66 à 0.90) par centimètre de taille supplémentaire
- A taille identique, un homme pèse en moyenne 3.2 kg (IC95% 0.9 à 5.5) de plus qu'une femme
- Ces 2 associations sont statistiquement significatives ($p<0.05$)



Changement moyen de poids par cm supplémentaire chez des personnes de même genre

Régression multiple: exemple Poids

- Le coefficient de détermination de ce modèle est 0.39: la taille <u>et</u> le genre explique 39% de la variance du poids

- Message: qualité de la prédiction améliorée par rapport au modèle simple Poids/Genre mais peu amélioré par rapport au modèle Poids/Taille

- Les pentes estimées ont des valeurs différentes car elles n'estiment pas les mêmes paramètres
  - modèles simples:
    - Poids prédit = – 86.3 + 0.88 * Taille
    - Poids prédit = 60.4 + 13.5 si Genre = Homme
  - modèle multiple:
    - Poids prédit = – 72.9 + 0.78 * Taille + 3.2 si Genre = Homme

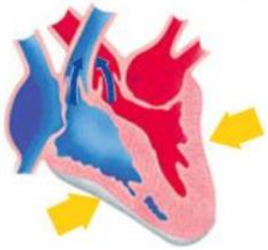
Sex-Specific Determinants of Left Ventricular Mass in Pre-Diabetic and Type 2 Diabetic Subjects

Diabetes Care 30:946–952, 2007

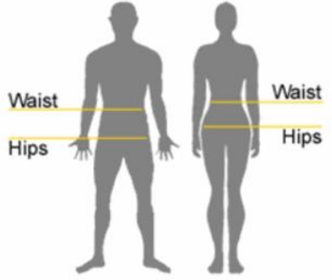
Bernhard Kuch, MD[1]
Wolfgang von Scheidt, MD[1]
Wolfgang Peter, MD[1]
Angela Döring, MD[2]
Wolfgang Piehlmeier, MD[3]
Rüdiger Landgraf, MD[3]
Christa Meisinger, MD, MPH[2,4]

- Variable dépendante: masse ventriculaire gauche
- Variables indépendantes:

Systolic blood pressure



waist-to-hip ratio (WHR)
(rapport taille-hanche)



Prise de médicaments pour l'hypertension (non/oui)

Sex-Specific Determinants of Left Ventricular Mass in Pre-Diabetic and Type 2 Diabetic Subjects

- Résultats du modèle de régression linéaire multiple

Coefficients de régression

	Women	
	β	P
Diabetic subjects without CVD		
Systolic blood pressure (mmHg)	0.60	0.003
WHR	274.6	0.007
Antihypertensives (yes/no)	25.6	0.02

Valeurs p des tests d'hypothèse nulle: coefficient = 0

- <u>A rapport taille-hanche et prise de médicaments pour l'hypertension identique</u>, pour chaque incrément de 1 mmHg de tension artérielle systolique, la masse ventriculaire gauche augmente en moyenne de 0.60 g
- <u>A tension artérielle systolique et rapport taille-hanche identiques</u>, une personne avec un traitement pour hypertension a une masse ventriculaire gauche plus grande de 25.6 g en moyenne qu'une personne sans traitement pour l'hypertension
- <u>A tension artérielle systolique et prise de médicaments pour l'hypertension identiques</u>, pour chaque incrément de 1 du rapport taille-hanche, la masse ventriculaire gauche augmente en moyenne de 274.6 g !
- Toutes ces associations sont statistiquement significatives

Sex-Specific Determinants of Left Ventricular Mass in Pre-Diabetic and Type 2 Diabetic Subjects

- Résultats du modèle de régression linéaire multiple

Coefficients de régression

	Women	
	β	P
Diabetic subjects without CVD		
Systolic blood pressure (mmHg)	0.60	0.003
WHR	274.6	0.007
Antihypertensives (yes/no)	25.6	0.02

Valeurs p des tests d'hypothèse nulle: coefficient = 0

- L'intercept (constante du modèle) n'était pas rapporté
- L'équation de régression est rapportée de manière incomplète => le lecteur ne peut pas calculer les valeurs de masse ventriculaire prédites par le modèle pour un de ses patients
- … mais la prédiction n'était pas l'objet de l'étude
- Le modèle de régression linéaire est utilisé ici pour détecter des associations

- Critique: les intervalles de confiance à 95% n'étaient pas rapportés

Messages importants

- Modèle linéaire simple (une variable indépendante):
  - droite de prédiction de la forme y = a + bx,
  - Y (variable dépendante) est quantitative continue
  - **a** et **b** sont estimations de paramètres (appelons les α et β) obtenues avec les données de l'échantillon:
    - intervalle de confiance à 95%
    - valeur p pour tester si α =0 ou β =0
  - a et b dépendent des unités de mesure

hypothèses nulles

- Attention: un modèle de régression linéaire n'est pas toujours approprié (relation non linéaire entre Y et X par exemple)

Messages importants

- Modèles multiples (plusieurs variables indépendantes)
  - interprétation des pentes différente de celle des modèles simples
  - pour améliorer la capacité du modèle à prédire Y
  - pour neutraliser des phénomènes de confusion: cours 13

- Modèles utiles pour détecter des associations et décrire l'influence entre la(les) variable(s) indépendante(s) et la variable dépendante

- Coefficient de corrélation
  - mesure le degré d'association linéaire entre 2 variables continues
  - X et Y peuvent représenter des quantités différentes
  - Coefficient de corrélation de Pearson $\boldsymbol{r}$
  - $\boldsymbol{r}^2$: pourcentage de variance de Y expliquée par X (et vice-versa)

Objectifs prochaine séance

- Comprendre les notions suivantes
  - principes de l'essai clinique randomisé
  - allocation aléatoire
  - placebo
  - insu
  - principaux biais

Chapitres Petrie/Sabin
13 – 14