Ecole Supérieure en Sciences Appliquées de Tlemcen

Analyse de données

Chapitre 3: Description bidimensionnelle et mesure de liaison entre variables

Partie 2

Présentée par:

Dr Imane NEDJAR

Introduction

L’objectif de la Statistique Descriptive est de décrire les données observées pour mieux les analyser

- **Liaison Entre Variables**
- **Régression linéaire**

Régression linéaire

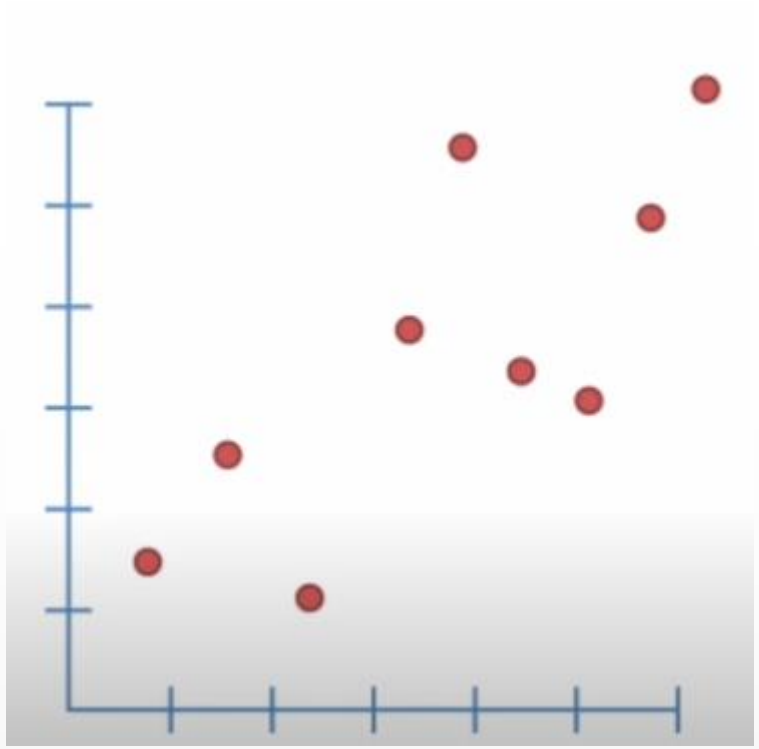
Régression linéaire

Lorsque deux variables quantitatives sont correctement corrélées
et que l'on peut considérer, a priori, que l'une (nous
supposerons qu'il s'agit de X) est cause de l'autre (il s'agira donc de Y)

➔ ***Chercher une fonction de X approchant Y « le mieux possible » en un certain sens***

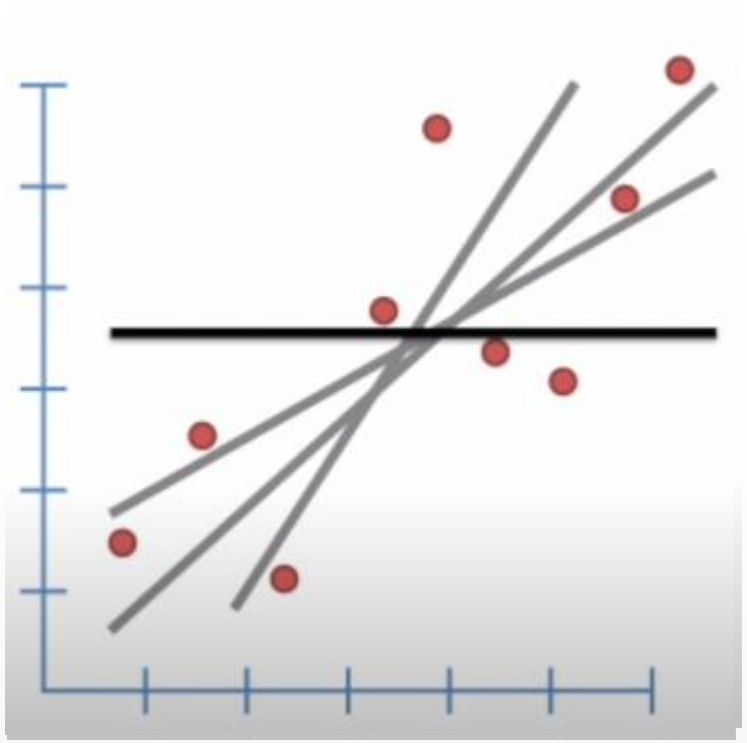
Régression linéaire

Poids



Taille

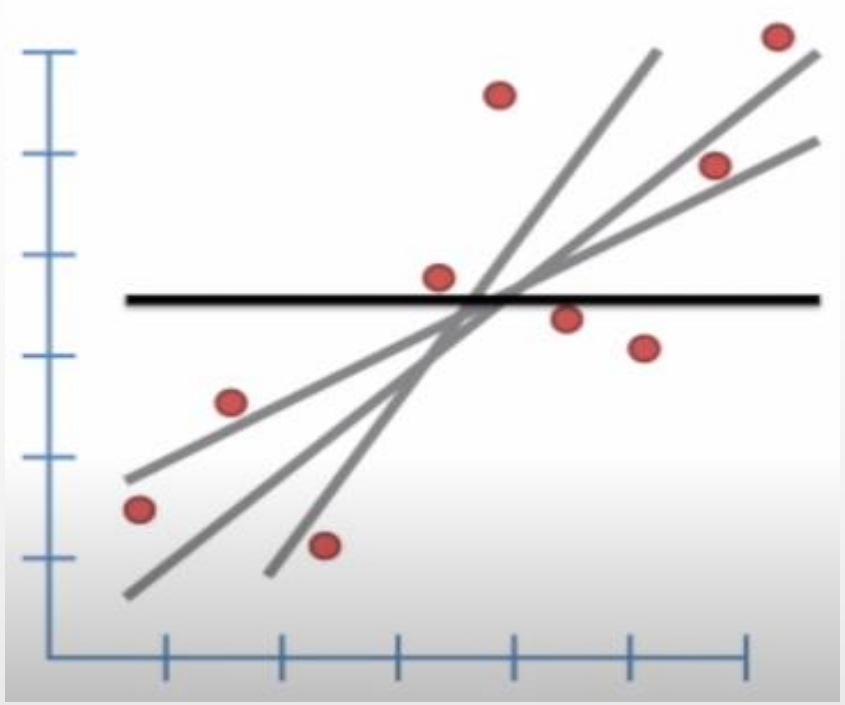
Poids



Taille

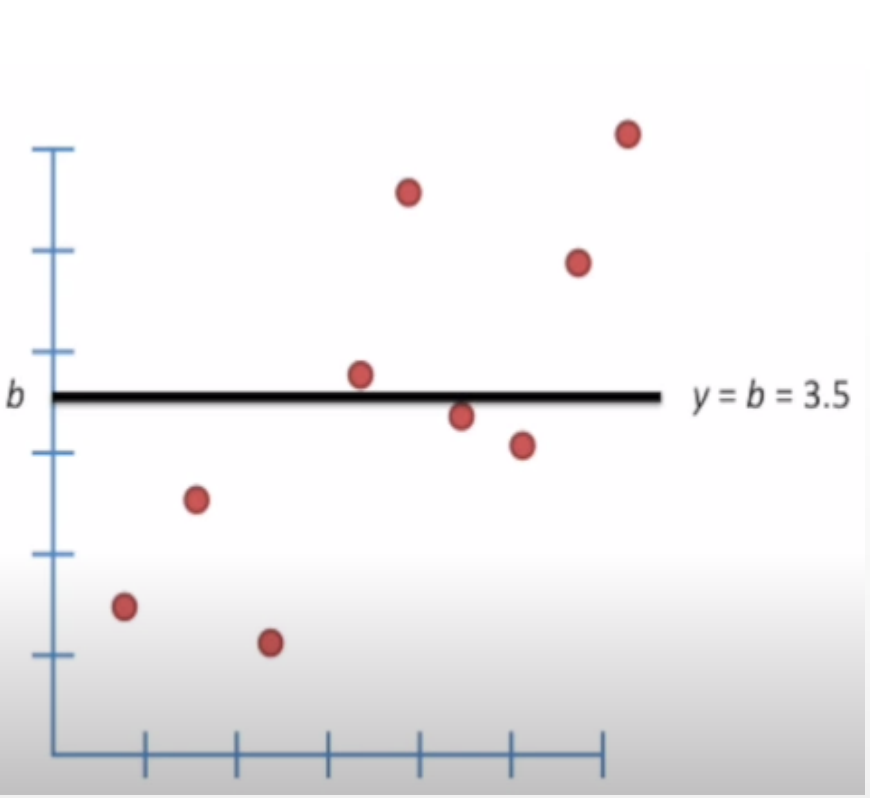
Régression linéaire

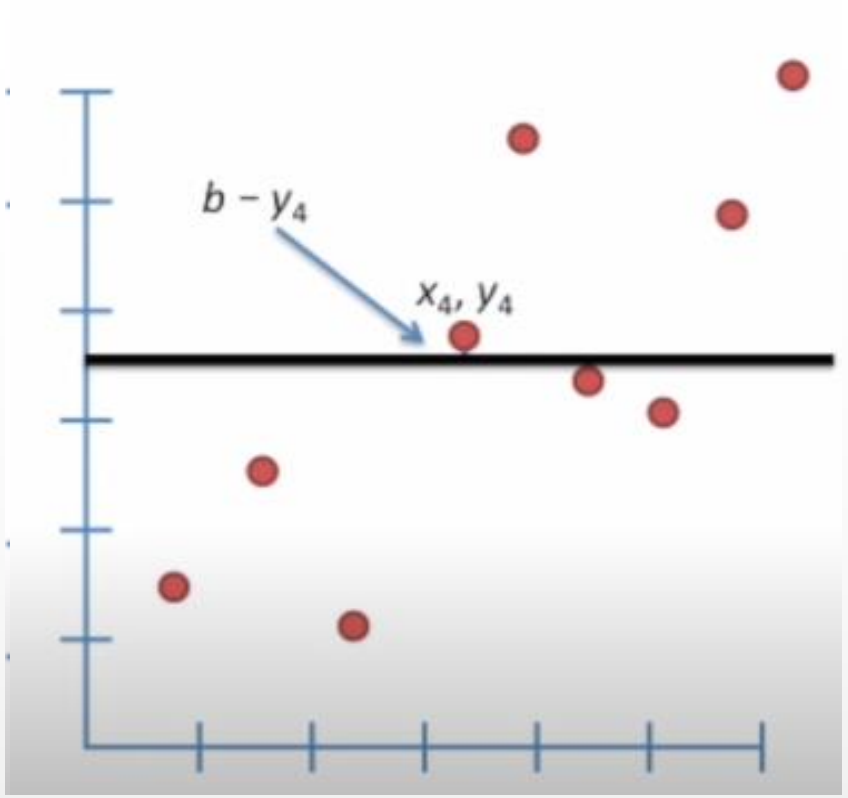
Régression linéaire tente de modéliser la relation entre deux variables en ajustant une équation linéaire aux données observées.

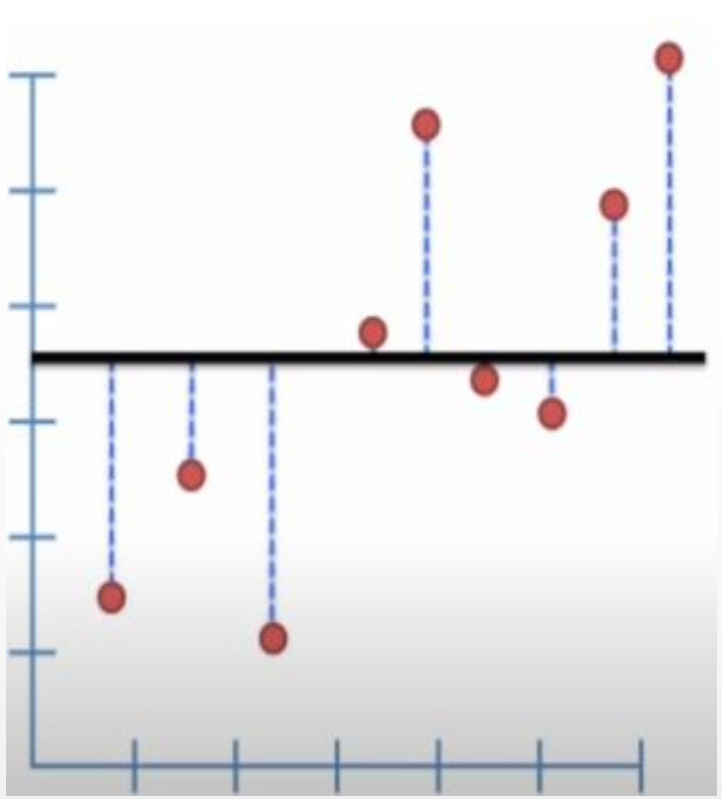


Régression linéaire

Nous pouvons mesurer l'adéquation de cette ligne aux données en voyant à quel point elle est proche des points de données





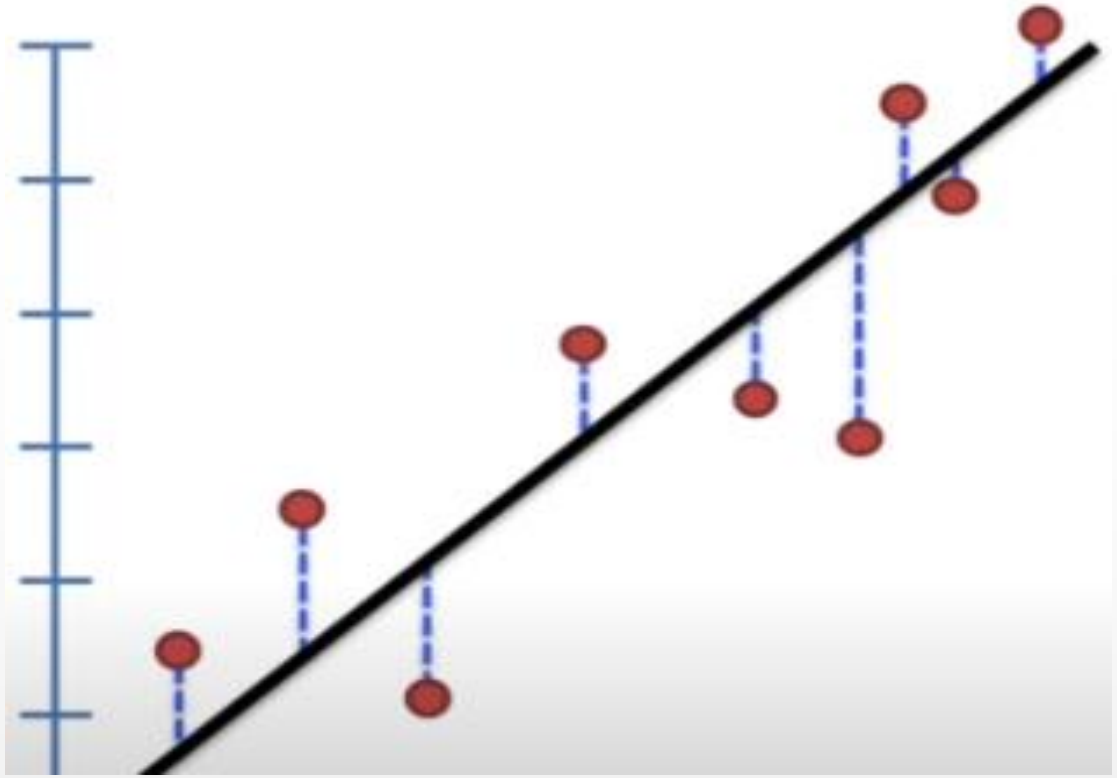


$$(b-y_1)^2+(b-y_2)^2+(b-y_3)^2+(b-y_4)^2+(b-y_5)^2+(b-y_6)^2+(b-y_7)^2+(b-y_8)^2+(b-y_9)^2 = 24{,}62$$

Somme des carrés des résidus

Régression linéaire

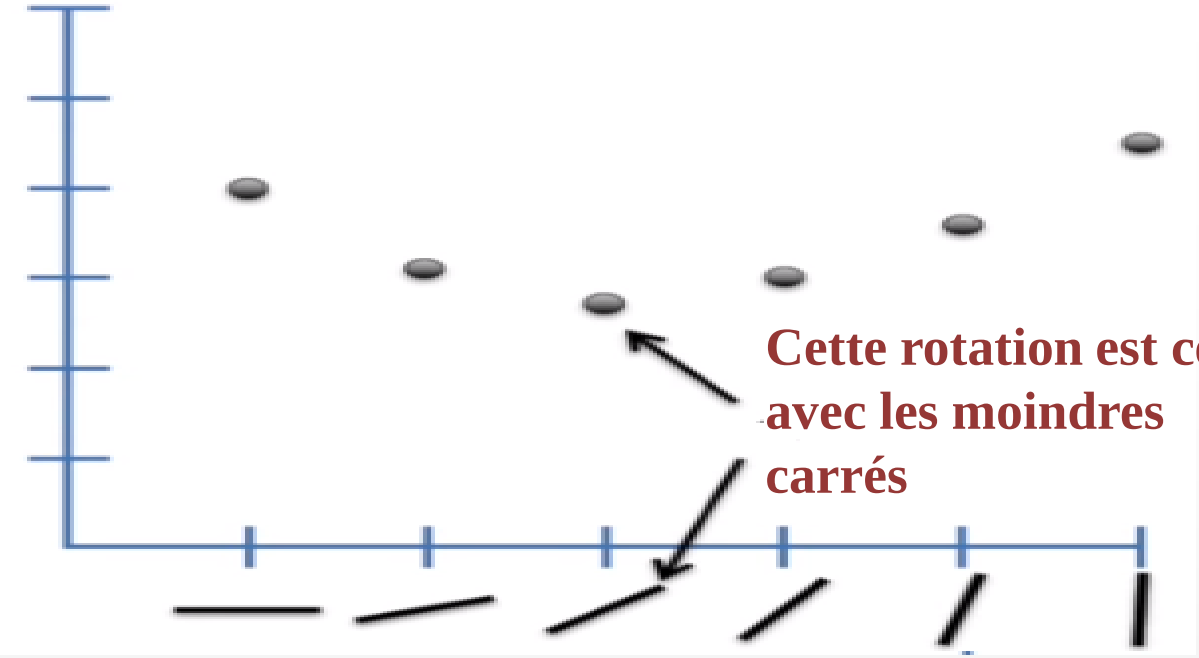
Tourner un peu la ligne



Avec la nouvelle ligne mesurez les résidus, mettez-les au carré puis additionnez les carrés

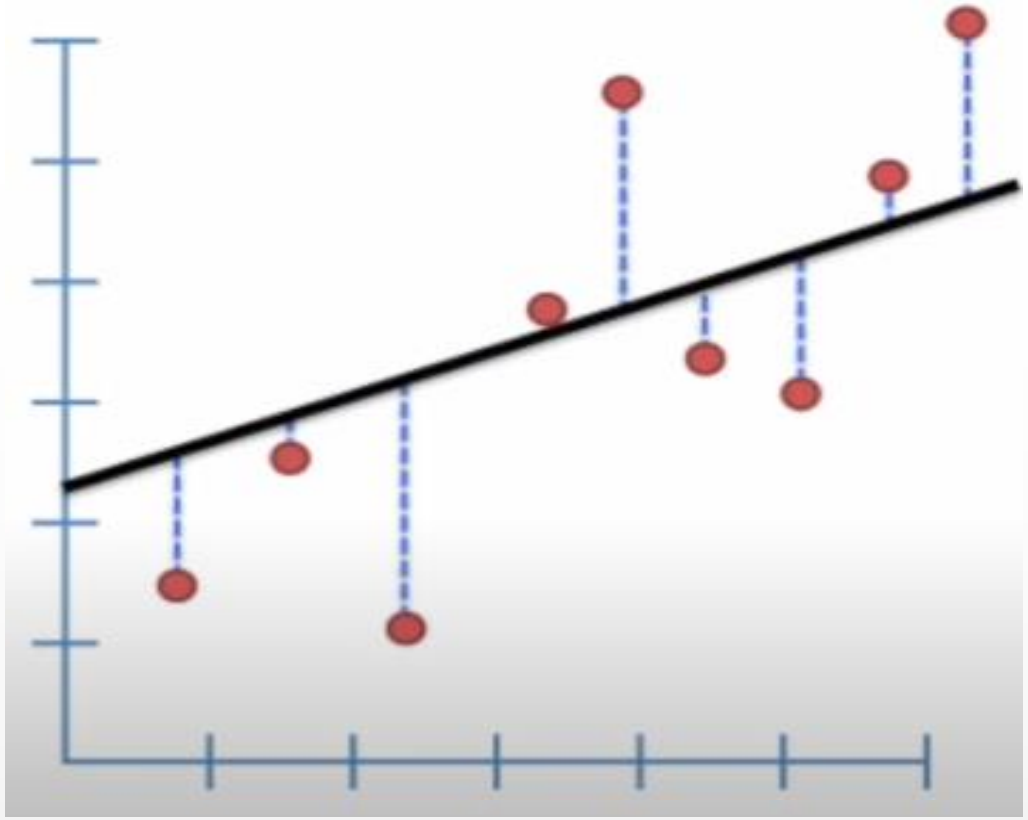
Régression linéaire

Somme des carrés des résidus



Différentes rotations

Régression linéaire



Somme des carrés des résidus

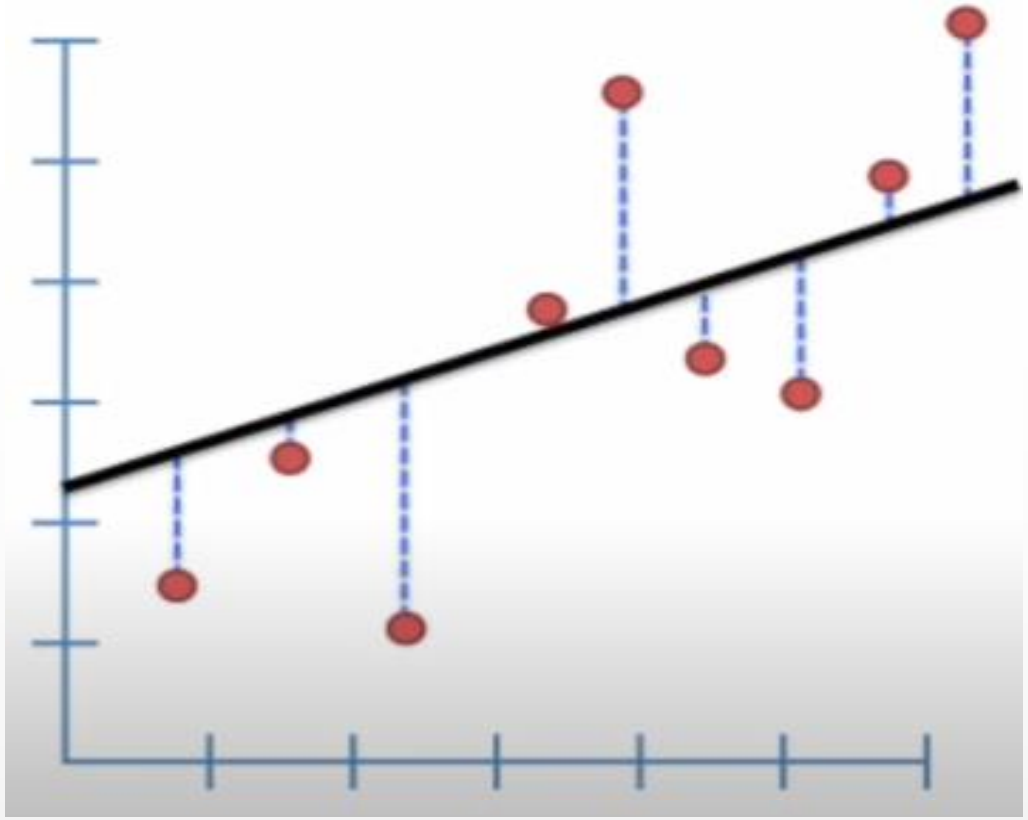
18.72

14.05

Y=ax+b

Nous voulons trouver des valeurs optimales pour a et b afin de minimiser la somme des carrés des résidus => Moindres carrés

Régression linéaire



$$Y = ax + b$$

$$a = \mathrm{cov}_{xy} / s_x^2$$

$$b = \bar{y} - a\,\bar{x}$$

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\mathrm{cov}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Nous voulons trouver des valeurs optimales pour a et b afin de minimiser la somme des carrés des résidus => Moindres carrés

Régression linéaire

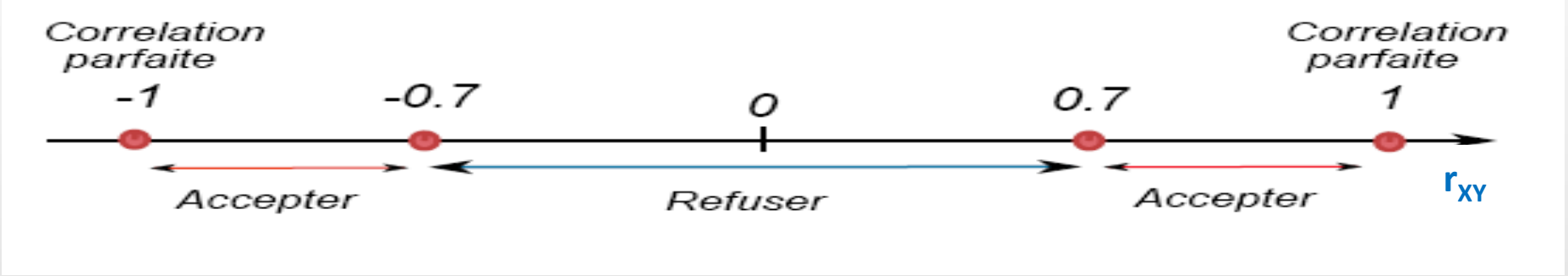
Evaluation de la qualité de la régression

Pour mesurer la qualité de l'approximation d'un nuage (xi, yi) i=1..n par sa droite des moindres carrés (après tout on peut toujours faire passer une droite par n'importe quel nuage !), on calcule son coefficient de corrélation linéaire défini par

$$r_{xy} = \mathrm{COV}_{xy} / s_x s_y$$

Si $|r_{XY}| < 0.7$ ➔ L'ajustement linéaire est refusé (Droite refusée).

Si $|r_{XY}| > 0.7$ ➔ L'ajustement linéaire est accepté (Droite acceptée).



Coefficient de détermination R-squared (R^2) -

- Est très similaire à son cousin R mais son interprétation est plus facile.
- Il est facile et intuitif de calculer

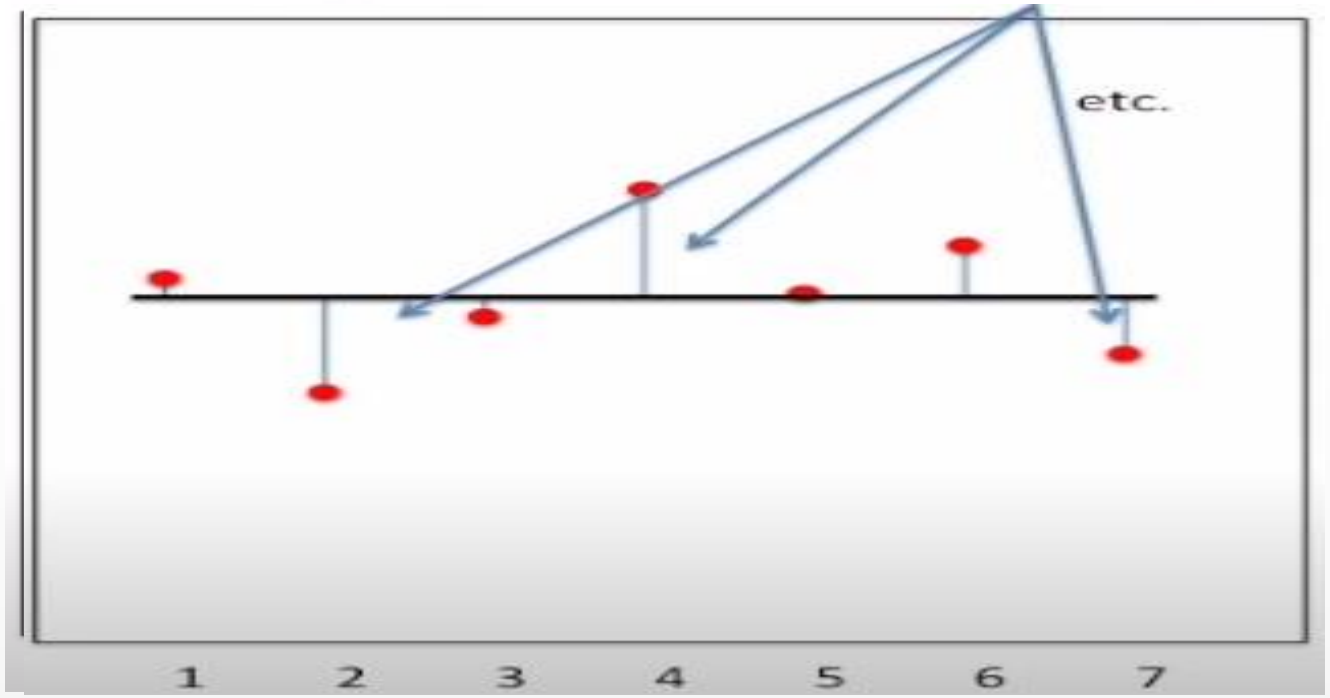
Coefficient de détermination R-squared (R^2) -

- Ce coefficient varie entre 0 et 1
- Permet de juger la qualité d'une régression linéaire.
- Si le $R^2=0$, cela signifie que l'équation de la droite de régression détermine 0 % de la distribution des points. Cela signifie que le modèle mathématique utilisé n'explique absolument pas la distribution des points.
- Si le R^2 vaut 1, cela signifie que l'équation de la droite de régression est capable de déterminer 100 % de la distribution des points.

R-squared (R^2)

La variation des données = somme (poids individuel i- moyenne)^2

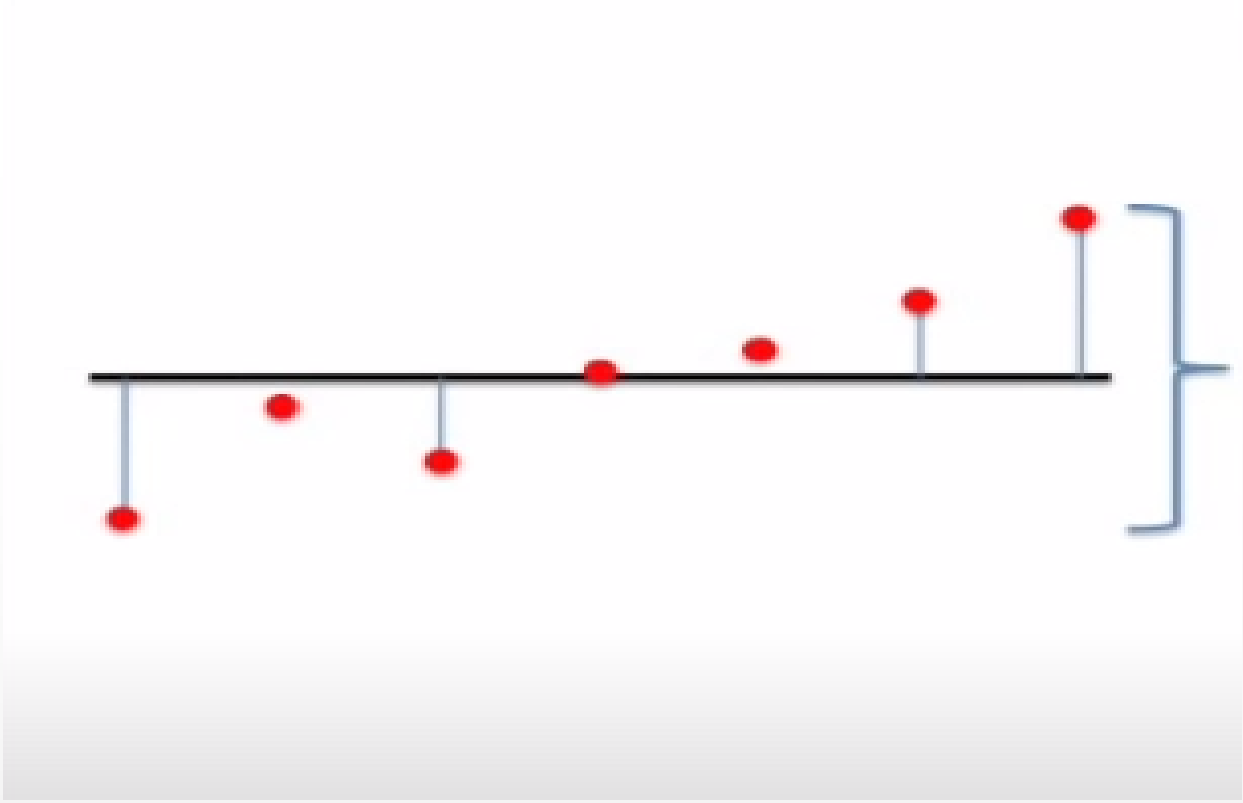
Poids



La moyenne Poids

#ID Individuel

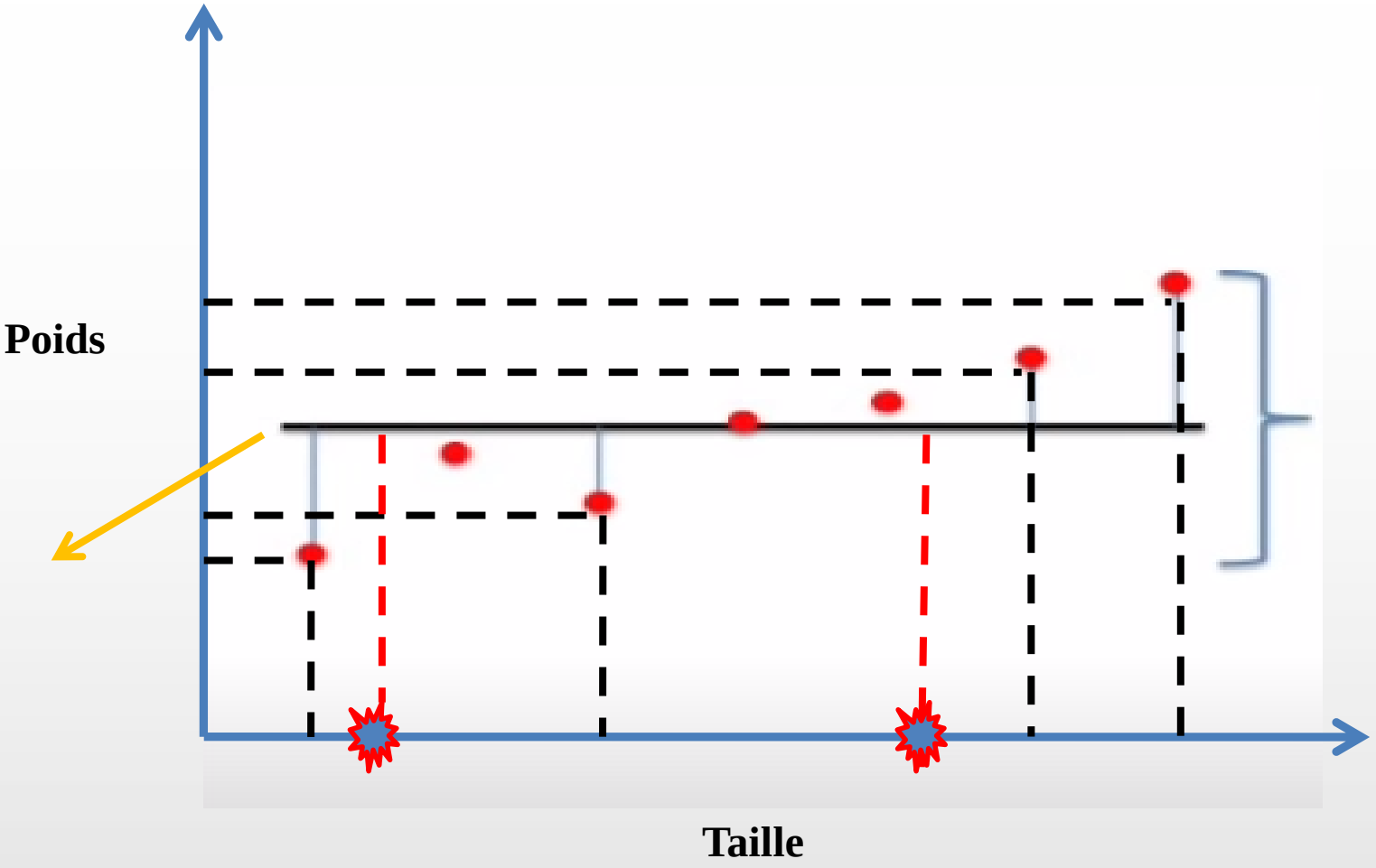
R-squared (R^2)



Tout ce que nous avons fait est de réorganiser les données sur l'axe X, la moyenne et la variation sont exactement les mêmes qu'avant

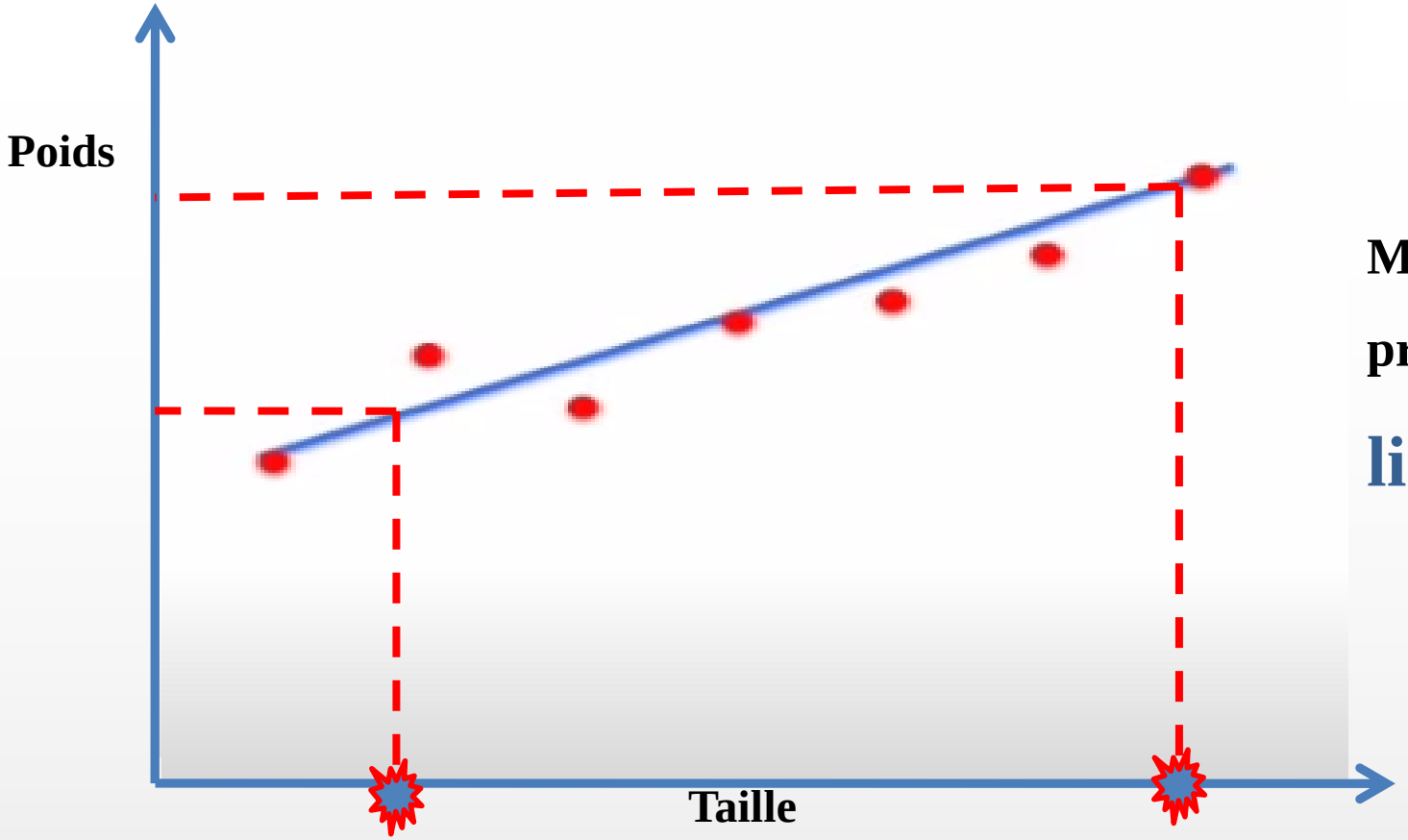
R-squared (R^2)

Etant donné que nous connaissons la taille d'un individu, la moyen est-il le meilleur moyen de prédire le poids des individus?



R-squared (R^2)

Ajuster une ligne sur les données

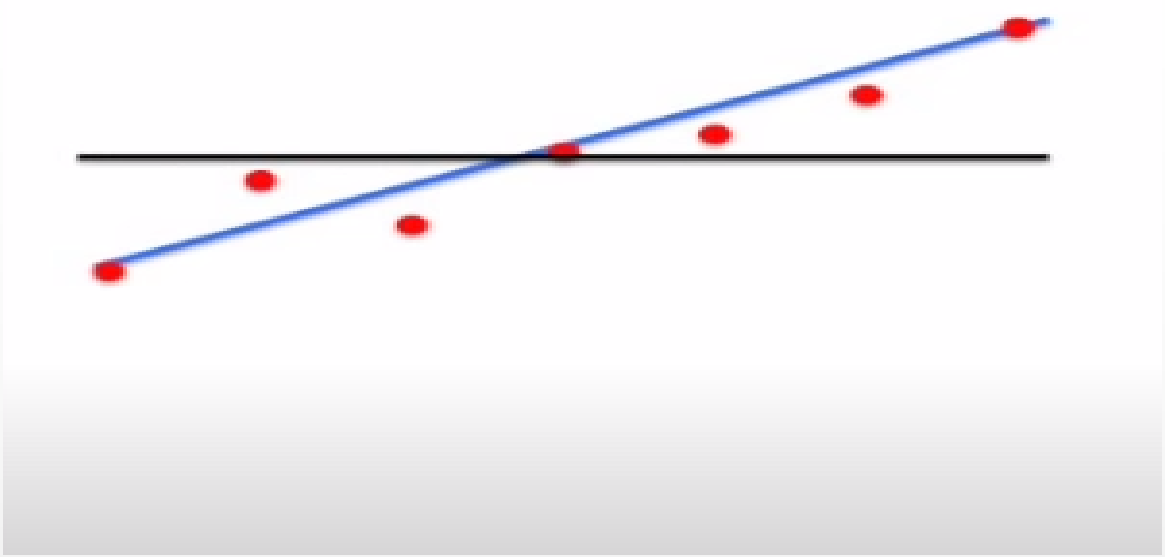


Maintenant, nous pouvons prédire le poids avec notre ligne

R-squared (R^2)

Question: la ligne décrit-elle mieux les données que la moyenne ?

Comment nous quantifions cette différence ?



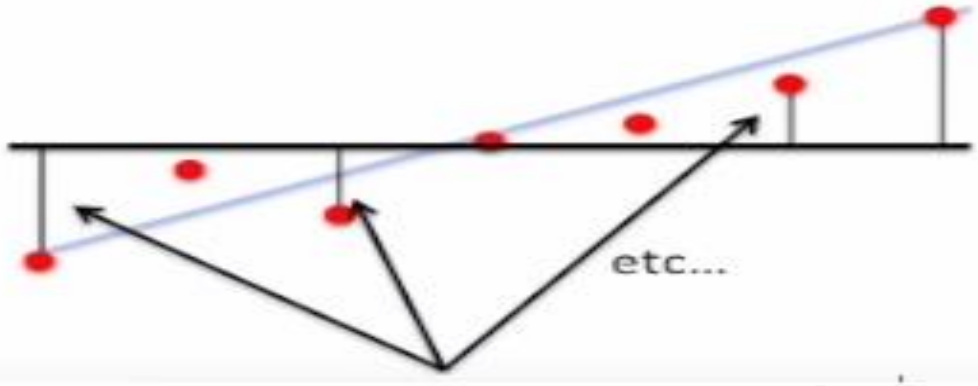
R^2

R-squared (R^2)

Quantifier la différence entre la droite et la moyenne

Calcule R^2

Poids



$$R^2 = \frac{Var(mean) - Var(line)}{Var(mean)}$$

Taille

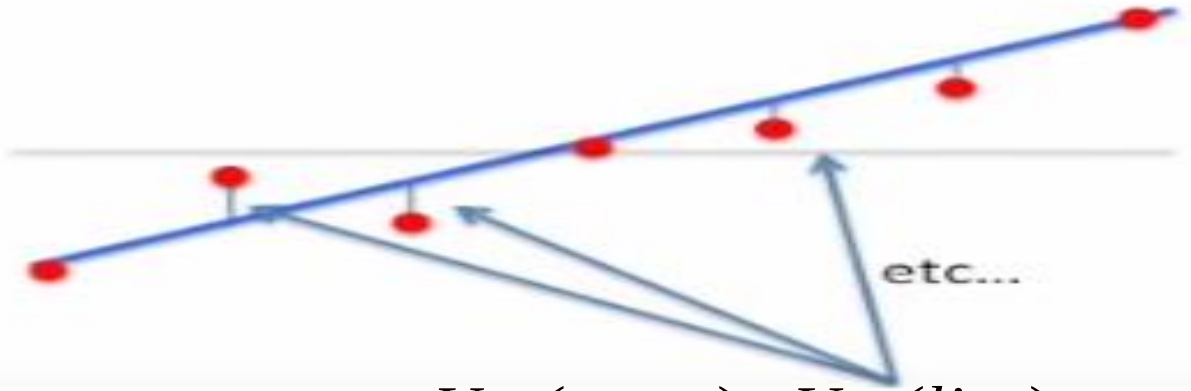
R-squared (R^2)

Quantifier la différence entre la droite et la moyenne

Calcule R^2

Var (mean) =32

Poids



$$R^2 = \frac{Var(mean) - Var(line)}{Var(mean)}$$

Taille

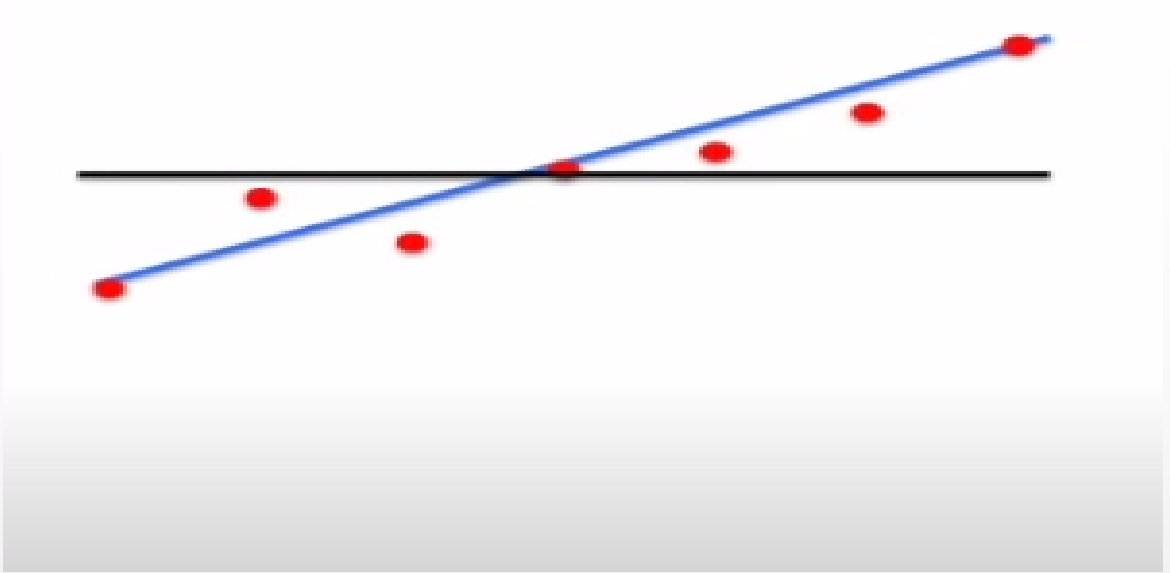
R-squared (R^2)

Quantifier la différence entre la droite et la moyenne

Calcule R^2

Var (mean) =32

Var (Line) =6



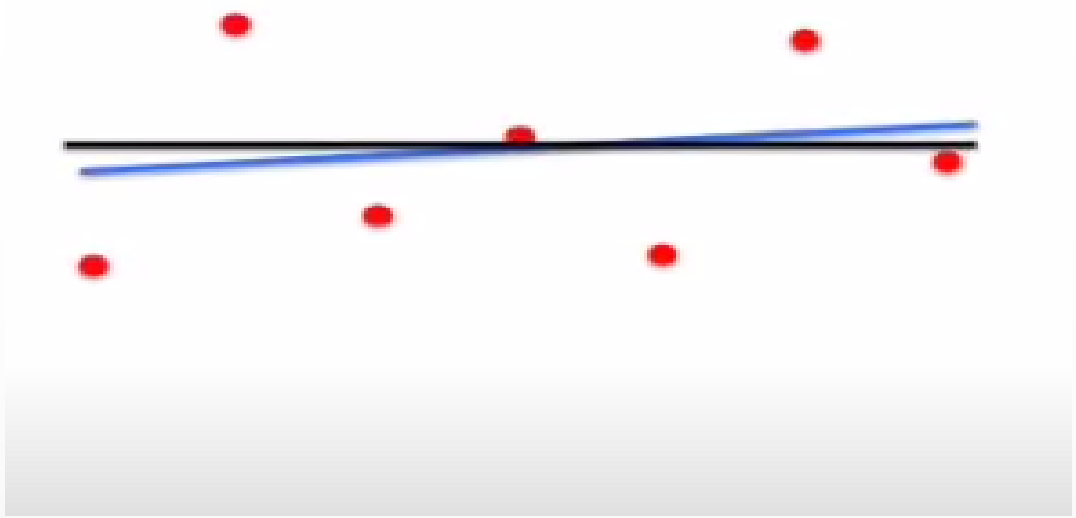
$$R^2 = \frac{32-6}{32} = 0.81 = 81\%$$

Il y a 81% moins de variation autour de la ligne que la moyenne

La relation poids/taille explique 81% de la variation des données

R-squared (R^2)

Exemple



Var (mean) =32

Var (Line) =30

$$R^2 = \frac{32-30}{32} = 0.06 = 6\%$$

Il n'y a que 6 % de variation en moins autour de la ligne que la moyenne

La relation Poids/ temps passé à travailler représente 6% de la variation

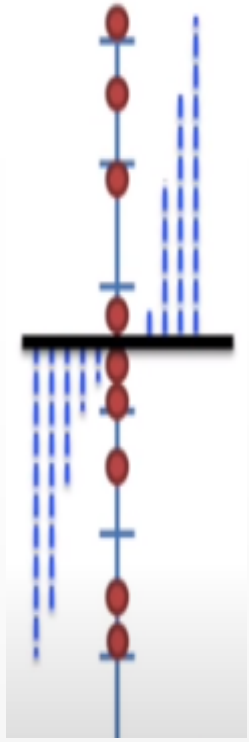
R-squared (R^2)

Exemple

La valeur de R^2 est de 0,9 => la relation entre les deux variables explique 90 de la variation des données

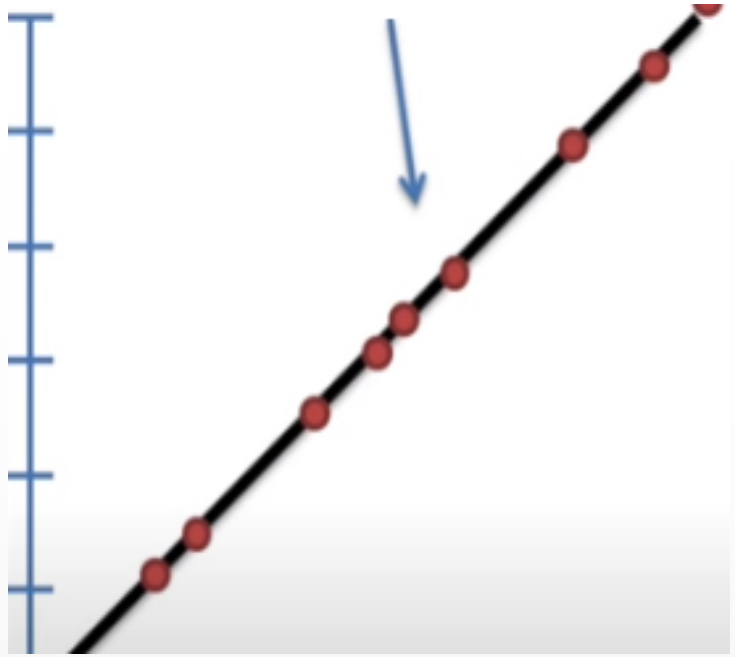
La valeur de R^2 est de 0,01 => il représente 1% de la variation des données, quelque chose d'autre doit expliquer le reste 99%

R-squared (R^2)



$$R^2 = \frac{Var(mean) - Var(line)}{Var(mean)}$$

$$R^2 = \frac{11.1 - 0}{11.1} = 1 = 100\%$$

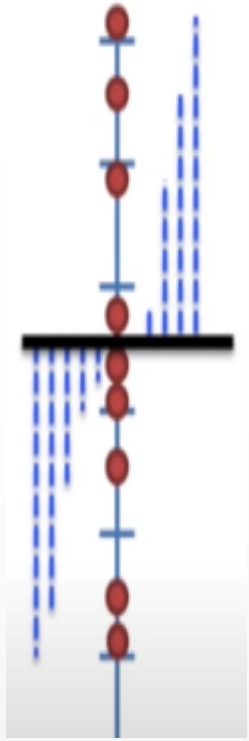


Var (mean)=11.1

Var (line)=0

Nous pouvons dire que la taille explique à 100% de la variation du poids

R-squared (R^2)

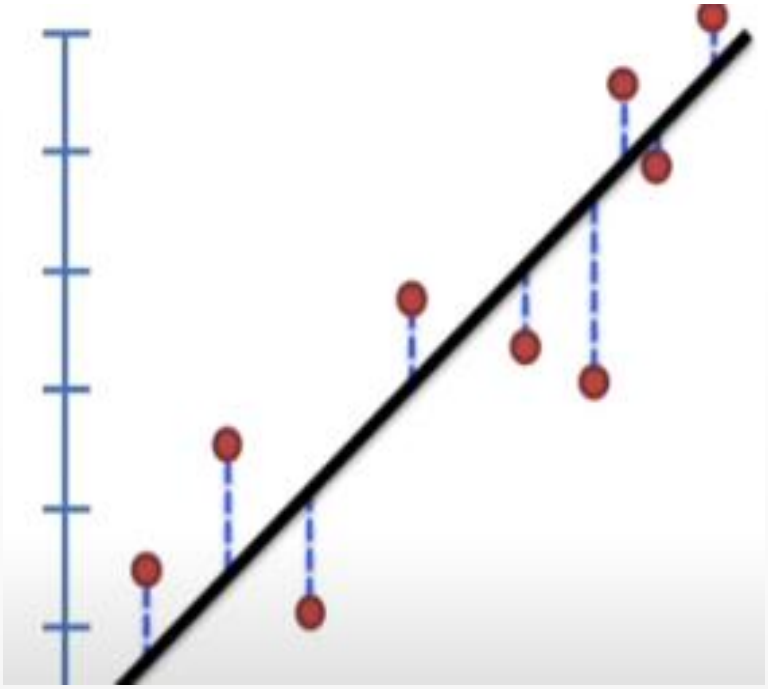


Var (mean)=11.1

R^2 Nous indique dans quelle mesure la variation de la taille de l'individu peut expliquée le poids

$$R^2 = \frac{Var(mean) - Var(line)}{Var(mean)}$$

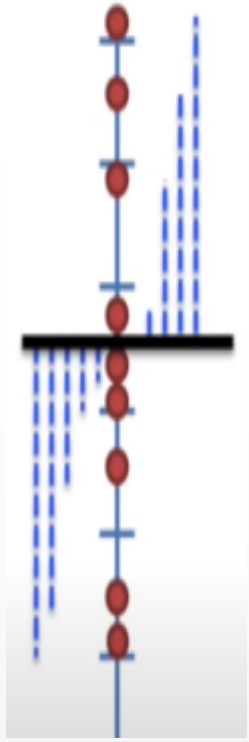
$$R^2 = \frac{11.1 - 4.4}{11.1} = 0.6 = 60\%$$



Var (line)=4.4

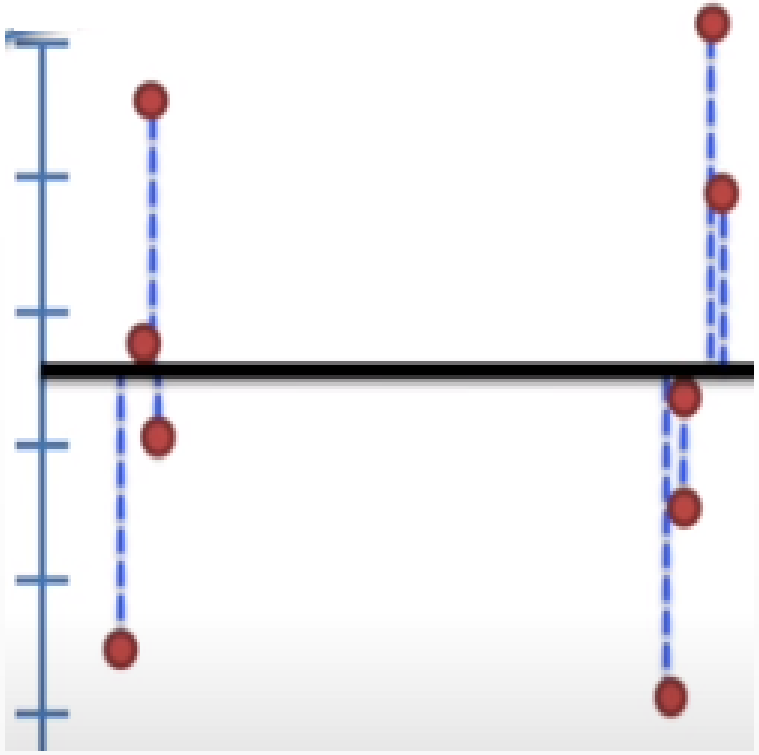
Nous pouvons dire que la taille explique 60% de la variation du poids

R-squared (R^2)



$$R^2 = \frac{Var(mean) - Var(line)}{Var(mean)}$$

$$R^2 = \frac{11.1 - 11.1}{11.1} = 0 = 0\%$$



Var (mean)=11.1

Var (line)=11.1

Nous pouvons dire que la taille n'explique aucune variation du poids

Coefficient de détermination R-squared (R^2)

- Est très similaire à son cousin R mais son interprétation est plus facile.
- Il est facile et intuitif de calculer

$R^2=0.7^2=0.5$ 50%

$R^2=0.5^2=0.25$ 25%

Avec R^2, il est facile de voir que la première corrélation est deux fois plus bonne que la seconde

Coefficient de détermination R-squared (R^2)

R^2 est le pourcentage de variation expliqué par la relation entre deux variables