Data exploration
Statistiques descriptives bivariées

- Observer simultanément des individus d'une population sur deux caractères
- Mesurer un lien éventuel entre deux caractères en utilisant un résumé chiffré qui traduit l'importance de ce lien
- Qualifier ce lien :
  - en cherchant une relation numérique approchée entre deux caractères quantitatifs
  - en cherchant des correspondances entre les modalités de deux caractères qualitatifs

2 types de variables $\Rightarrow$ 3 types de croisements :
- qualitatif $\times$qualitatif
- qualitatif $\times$quantitatif
- quantitatif $\times$quantitatif

Croisement Quantitatif - Quantitatif
Nuage de points

On considère X et Y deux variables quantitatives sur un échantillon de taille n. Les objectifs sont :

- Déterminer s'il y a un lien (corrélation) entre les deux variables.
- Construire un modèle permettant d'expliquer Y par X (ou vice-versa) s'il y a un lien.

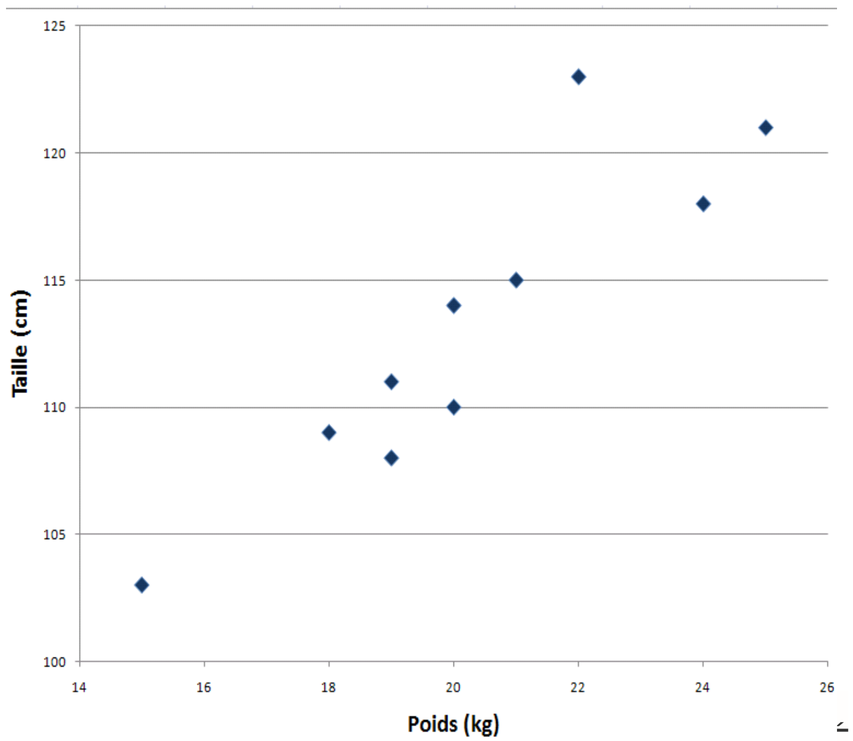
Le modèle pourra alors servir à faire de la prévision, c-a-d prévoir des valeurs de Y pour de nouvelles valeurs de X.

Etude du lien entre la taille (X) et le poids (Y) chez les enfants de 6 ans

Enfant	1	2	3	4	5	6	7	8	9	10
Taille	**121**	**123**	**108**	**118**	**111**	**109**	**114**	**103**	**110**	**115**
Poids	**25**	**22**	**19**	**24**	**19**	**18**	**20**	**15**	**20**	**21**

La première étape consiste à constater visuellement si ce lien existe. La représentation graphique appropriée est le *nuage de points.*

On cherche à repérer une forme particulière dans le nuage qui traduirai le lien entre X et Y. En particulier, une forme allongée traduit une relation de droite entre les deux variables.



Croisement Quantitatif - Quantitatif

Comment quantifier le lien entre X et Y?

La *covariance empirique* est un indicateur numérique du lien entre X et Y,

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

En développant la somme, on obtient

$$c_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

$$C_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y}) = \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \bar{x}\underbrace{\frac{1}{n}\sum_{k=1}^{n} y_k}_{\bar{y}} - \bar{y}\underbrace{\frac{1}{n}\sum_{k=1}^{n} x_k}_{\bar{x}} + \bar{x}\,\bar{y}$$

$$= \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \bar{x}\,\bar{y} - \bar{y}\,\bar{x} + \bar{x}\,\bar{y} = \frac{1}{n}\sum_{k=1}^{n} x_k y_k - \bar{x}\,\bar{y}$$

Plus elle est éloignée de 0, plus les variables sont liées. L'inconvénient est qu'elle n'est par normée. Pour pallier ce problème, on définit le *coefficient de corrélation linéaire* (coefficient de Pearson) à valeurs dans [-1,1]

$$r_{xy} = \frac{c_{xy}}{s_x\, s_y}.$$

Si le coefficient de corrélation linéaire est proche de 1 en valeur absolue, alors un modèle de type équation de droite est possible entre X et Y.

Rq. Si les séries sont réduites ($s_x^2 = s_y^2 = 1$), le coefficient de corrélation correspond à la covariance

Croisement Quantitatif - Quantitatif
Droite de régression

On note $\{x_i\}_{i=1,\dots,n}$ la série observée pour X et $\{y_i\}_{i=1,\dots,n}$ la série observée pour Y.

L'objectif est de trouver une fonction *f* telle que

$$y_i = f(x_i) + \varepsilon_i$$

où ε représente l'erreur.

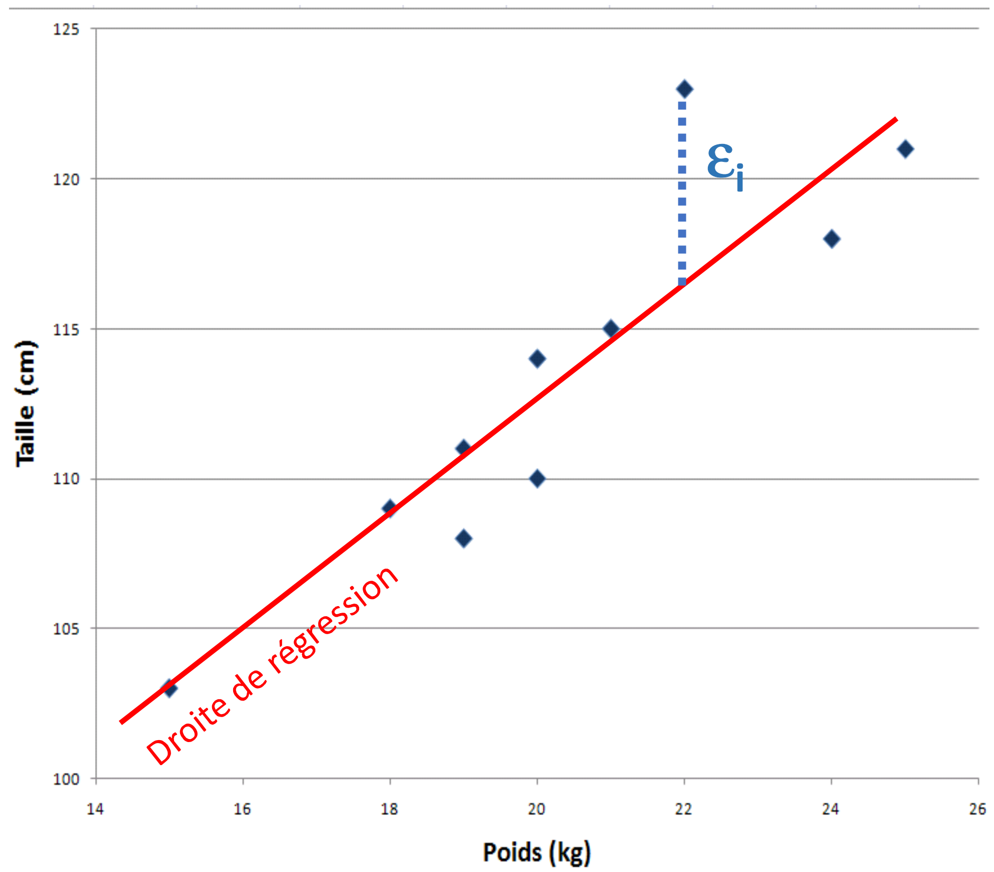
On se restreint aux fonctions affines :

$$f(x) = ax + b$$

Et on cherche les coefficients a et b qui minimisent *l'erreur quadratique moyenne*

$$EQ(a,b) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - (ax_i + b))^2$$



Rq. L'erreur quadratique moyenne est aussi appelée fonction de coût en *machine learning* et peut prendre d'autres formes.

Croisement Quantitatif - Quantitatif

Coefficients de la droite de régression

Par minimisation de l'erreur quadratique moyenne,

$$\frac{\partial}{\partial a} EQ(a,b) = 0 \quad \text{et} \quad \frac{\partial}{\partial b} EQ(a,b) = 0,$$

on obtient les coefficients :

$$\hat{a} = \frac{c_{xy}}{s_x^2} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

Le « chapeau » au dessus de a et b signifie que la valeur obtenue est une estimation sur un échantillon

$y = \hat{a}x + \hat{b}$ est appelée *droite de régression* de Y en X. Elle traduit les variations de Y qui peuvent être expliquées par X. Attention la droite de régression de X en Y n'est nécessairement la même que celle de Y en X

Exemple : Etude du lien entre l'âge et le poids chez les enfants de 6 ans

$\bar{x}$	$\bar{y}$	s_x^2	s_y^2	c_{xy}
113,20	**20,30**	**38,62**	**8,46**	**16,27**

L'équation de la droite de Y en X : y=0,42 x – 27,38

L'équation de la droite de X en Y : y=1,92 x – 74,15

Croisement Quantitatif - Quantitatif

Lien entre pente de la droite et coefficient de corrélation

On a la pente de la droite qui est proportionnelle au coefficient de corrélation

$$\hat{a} = \frac{c_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x}.$$

On peut donc faire les interprétations suivantes :

- $|r|$ est proche de 1 alors X et Y sont très liés entre eux par une droite affine.
- $r < 0$: globalement X et Y varient en sens inverse .
- $r > 0$: globalement X et Y varient dans le même sens .
- $|r| \cong 0$: on ne peut rien dire sur un lien éventuel entre X et Y.

$|r| \cong 0$ ne signifie pas qu'il n'y a pas de lien entre X et Y mais uniquement que le lien n'est pas linéaire. Nous verrons en TD que le coefficient de corrélation correspond au cosinus de l'angle entre les deux vecteurs X et Y.

Exemple : Etude du lien entre l'âge et le poids chez les enfants de 6 ans

On trouve

$$\mathbf{r_{xy}=0,90}$$

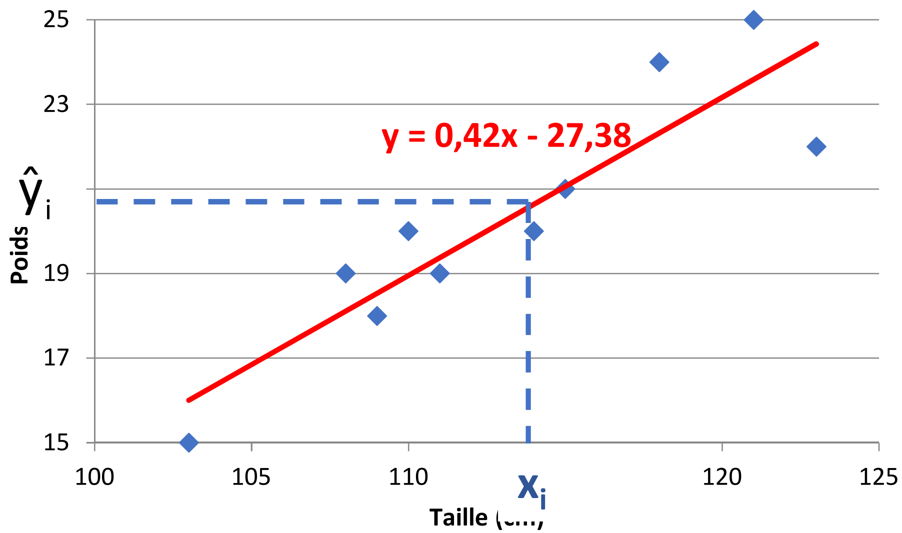
- $r_{xy} \cong 1 \Rightarrow$ L'équation de droite est donc pleinement justifiée
- $r_{xy} > 0 \Rightarrow$ plus la taille est grande et plus le poids est important (et vice-versa)

Croisement Quantitatif - Quantitatif

Prévisions

On appelle *prévisions* les valeurs données par la droite de régression. Pour chaque point x_i de la série observée, on peut calculer la prévision (*i.e.* une valeur approchée de y_i par la droite de régression)

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$



<u>Propriétés</u> :

La variable Y et la partie de cette variable expliquée par la droite de régression ont la même moyenne :

$$\bar{\hat{y}} = \bar{y}$$

Démonstration en TD

mais pas la même variance :

$$s_{\hat{y}}^2 = s_{\hat{y}}^2 \times r_{xy}^2$$

$$\hat{y} = \hat{a}x + \hat{b} \quad \Rightarrow \quad s_{\hat{y}}^2 = (\hat{a})^2 s_x^2 = \left(r_{xy}\frac{s_y}{s_x}\right)^2 s_x^2 = r_{xy} s_y^2$$

$\Longrightarrow$ La variance de Y expliquée la droite de régression est plus petite que la variance de Y

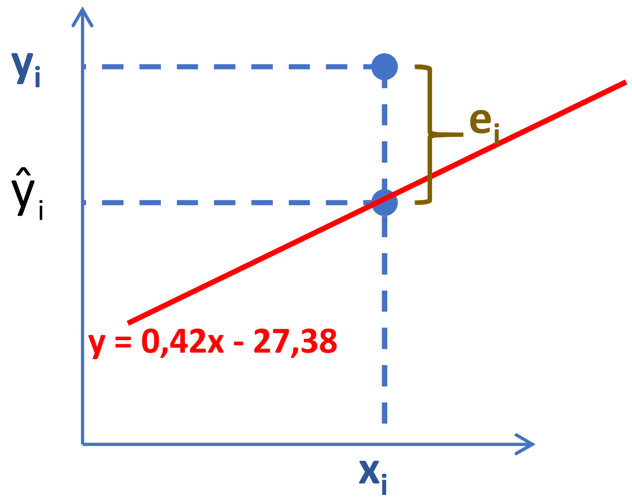
$\Longrightarrow$ La variance de Y expliquée la droite de régression est d'autant meilleure que le coefficient de Pearson est proche de 1 en valeur absolue.

Croisement Quantitatif - Quantitatif

Résidus

On appelle *résidus* l'écart entre la valeur observée y_i et la valeur prédite $\hat{y}_i$

$$e_i = y_i - \hat{y}_i = y_i - (\hat{a}x_i + \hat{b})$$



Démonstration en TD

La moyenne des résidus est nulle : $\bar{e} = 0$ donc l'erreur globale est égale à la variance des résidus : $EQ\left(\hat{a}, \hat{b}\right) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = s_e^2$

Et on peut montrer que

$$s_e^2 = s_y^2\left(1 - r_{xy}^2\right)$$

$$EQ = s_e^2 = s^2_{\{y-(\hat{a}x+\hat{b})\}} = s^2_{y-\hat{a}x} = s_y^2 + (\hat{a})^2 s_x^2 - 2\hat{a}C_{xy} = s_y^2 + \left(r_{xy}\frac{s_y}{s_x}\right)^2 s_x^2 - 2\left(r_{xy}\frac{s_y}{s_x}\right)(s_x s_y r_{xy}) = s_y^2(1 - r_{xy}^2)$$

$\Rightarrow$ L'erreur globale/variance des résidus est proportionnelle à la variance de la variable Y

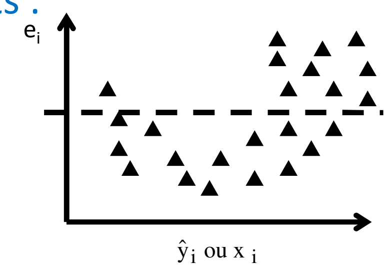
$\Rightarrow$ L'erreur est d'autant plus petite que le coefficient est proche de 1 en valeur absolue

Validité du modèle :

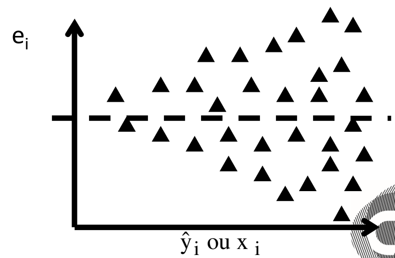
Un modèle est explicatif s'il ne reste plus « d'information » dans les résidus pouvant expliquée y.

On vérifie (graphiquement) les trois points suivants :

- La moyenne des résidus est nulle
- Les résidus ne sont pas corrélés
- La variance des résidus est constante



Résidus corrélés



Variance non constante

Croisement Quantitatif - Quantitatif

Décomposition de la variance

Nous avons vu que la variance de la variable Y n'est pas égale à la variance des valeurs prédites. Cependant elle peut se décomposer comme suit :

$$\underbrace{s_y^2}_{\text{variance totale}} = \underbrace{s_{\hat{y}}^2}_{\text{variance expliquée}} + \underbrace{s_e^2}_{\text{variance résiduelle}}$$

Démonstration en TD

En divisant cette égalité par la variance totale, on obtient le pourcentage de variance de y expliquée par le modèle, ce qu'on appelle encore le coefficient de détermination,

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = r_{xy}^2 \in [0,1]$$

Dans l'exemple précédent, on a la décomposition de la variance suivante :

	Variances
Régression	6,17
Résidus	1,44
Total	7,61

D'où R^2=6,17/7,61=0,81. Cela signifie que 81% de la variation des poids observés est expliquée par la droite de régression : poids = 0,42×taile - 27,38

Croisement Quantitatif - Quantitatif

Outliers

Un modèle peut s'avérer très précis pour ajuster les valeurs observées mais très mauvais en ce qui concerne la prévision de nouvelles valeurs.

Observation est *influente* si une faible variation entraine une modification importante des caractéristiques du modèle.

Détection des observations influentes (*atypiques/outliers*)

- On retire la i^ème^ observation de l'ensemble des données
- On ajuste un nouveau modèle sans la i^ème^ donnée
- On calcule $y_{(-i)}$ la prévision de y_i avec le nouveau modèle
- On calcule le résidus , $e_{(-i)}=y_i-y_{(-i)}$

✓**Un résidus important signale une observation influente**

On a $e_{(-i)} = \dfrac{e_i}{1-h_{ii}}$ où $h_{ii} = \dfrac{1}{n} + \dfrac{1}{(n-1)} \dfrac{(x_i - \bar{x})^2}{s_x^2}$

Un *levier*

$1/n \le h_{ii} \le 1$

proche de 1 indique une observation influente

Le PRESS *(predicted residual sum of squares)* donne une indication sur les qualités prédictives du modèle

$$PRESS = \frac{1}{n}\sum_{i=1}^{n} e_{(-i)}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{e_i^2}{(1-h_{ii})^2}$$

Sous l'hypothèse de normalité des résidus, les *résidus studentisés*,

$$\delta_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

doivent être compris (IDC) entre ± 2

Croisement Quantitatif - Quantitatif
Transformation

- Les droites de régression n'explique que les liaisons linéaires.
- Si *X* et *Y* sont liées par une relation de la forme $Y=aX^2$ alors $r_{XY}=0$
  Le coefficient de corrélation linéaire de Pearson ne peut pas détecter cette liaison.
- Il n'existe pas de mesure universelle pour détecter des relations quelconques
- On essaie par des transformations de se ramener à une droite affine

Famille	Fonctions	Transformation	Forme affine
exponentielle	$y = a.e^{bx}$	$y' = \log(y)$	$y' = \log(a) + b.x$
puissance	$y = ax^b$	$y' = \log(y)\ x' = \log(x)$	$y' = \log(a) + b.x'$
inverse	$y = a + \frac{b}{x}$	$x' = \frac{1}{x}$	$y' = a + b.x'$
logistique	$y = \frac{1}{1 + e^{-(a.x+b)}}$	$y' = \log\left(\frac{y}{1-y}\right)$	$y' = a.x + b$

Croisement Quantitatif - Quantitatif
Méthodologie

1. Etude du nuage de points
2. Quantification du lien (coefficient de corrélation)
3. Construction du modèle
4. Détection des *outliers*
    - Etude des résidus standardisés : détection de potentiels *outliers*
    - Confirmation d'un *outlier* : retirer l'observation de l'étude, réajuster le modèle si changement alors *outlier* confirmé
5. Vérification des hypothèses sur les résidus
    - Si résidus corrélés, envisager une transformation des variables
6. Pourcentage de variabilité de Y expliquée par le modèle (R^2)
7. Utilisation du modèle pour prévision