

Statistiques pour la SAé 2.04

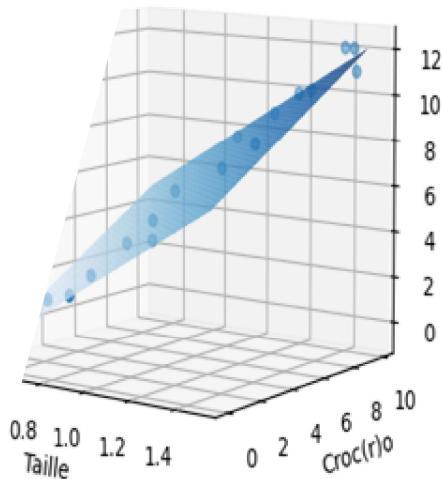
Cours 1 + TP1 + TP2

Régression linéaire multiple

S2.04 - Exploitation d'une Base de Données

R2.08 - Statistiques

2024-2025



Plan du cours

1. L'Analyse de données, c'est quoi ?
2. Comment ça, plusieurs dimensions ?
3. Normalisation des données : centrer, réduire
4. La régression linéaire multiple
5. Analyse des paramètres obtenus
6. Coefficient de corrélation multiple

• La SAé 2.04 : Exploitation d'une Base de Données

Données de la SAé 2.04

Travail sur des données sur les étudiant.es de DUT Informatique de Lannion entre 2016 et 2020 :

- Notes dans les différentes ressources
- Etablissement et ville d'origine
- Type de Bac
- etc.

Étapes de la SAé 2.04

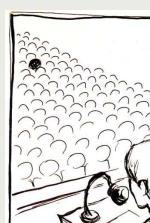
Le travail se fait en 3 parties successives, chacune avec un rendu à déposer sur Moodle :

- Partie 1 - Traduction et implantation d'un schéma relationnel
- Partie 2 - Peuplement et exploitation de la base de données
- **Partie 3 - Analyse statistique des données**

• Les Statistiques dans la SAé 2.04

Organisation des Statistiques pour la SAé 2.04

- Pendant 4 semaines, chaque semaine :
 - ≈ 2h TD sur PC (Python)
avec T. Jézéquel (A,B,C) ou Jean-Baptiste Faure (D et E)
- ↪ apprentissage d'outils qui seront à utiliser sur la Partie 3 (Statistique) de la SAé
- Puis, pendant la dernière semaine :
 - 1h de projet encadré (par T.Jézéquel ou J-B. Faure)
 - 1h de projet encadré (par enseignant.e de BD)
 - 2h en autonomie
- ↪ travail sur la Partie 3 (Statistique) de la SAé.



Évaluation de la Partie 3 (Statistique)

- Un rendu sous forme de rapport d'environ 6 à 10 pages.
- 1 note sur ce rapport, qui compte autant que celle d'un rendu de BD.



1. L'Analyse de données, c'est quoi ?

Dans cette SAé : premiers pas en **Science des données**.

Définition

La **science des données** (*data science* en anglais) est un domaine interdisciplinaire qui utilise les mathématiques, les statistiques, le calcul scientifique, les méthodes scientifiques, les process, les algorithmes et les systèmes informatiques automatisés pour extraire et extrapoler des connaissances à partir de grandes quantités de données brutes.



• Analyse de données

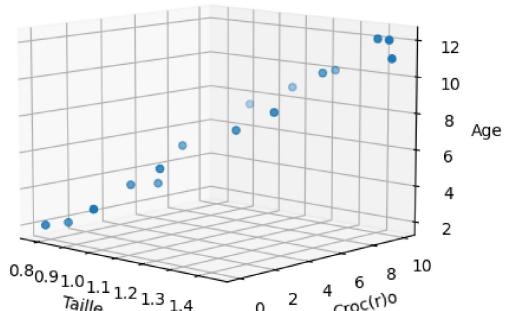
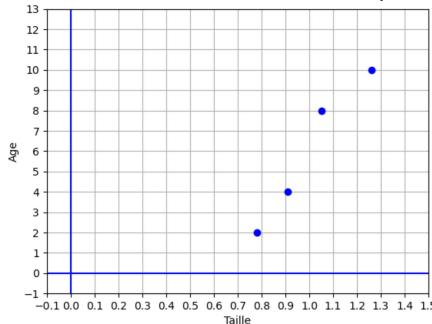
Dans la partie Statistique de la SAé, on sera plus particulièrement dans le domaine de l'**analyse de données**.

Définition

L'**analyse des données** est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives.

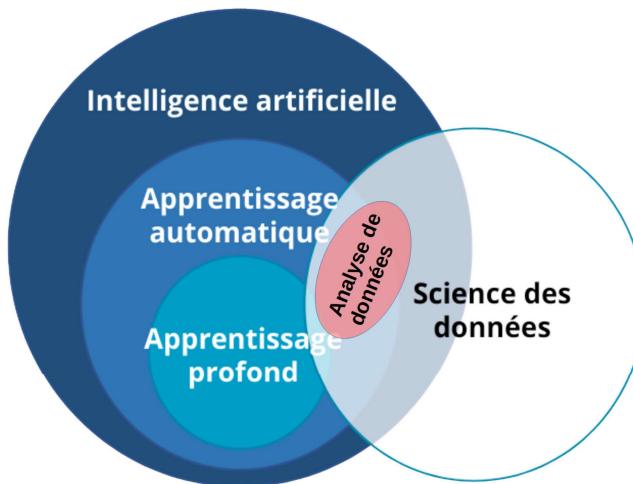
"multidimensionnelles" : dans le Cours 3 on a vu des représentations graphiques de 2 variables statistiques... donc 2 dimensions.

Ici, on va étudier des cas à plus de 2 variables !



• Méthodes d'analyse de données

Il y a 3 principales catégories de méthodes d'analyse de données, qui se trouvent être les 3 catégories principales d'Apprentissage automatique (Machine Learning en anglais).



Ainsi, on va retrouver ces 3 grandes catégories de méthodes dans la bibliothèque Python de Machine Learning Scikitlearn

scikit-learn

Machine Learning in Python

Getting Started

Release Highlights for 1.4

[GitHub](#)

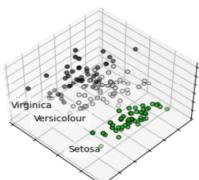
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: PCA, feature selection, non-negative matrix factorization, and more...

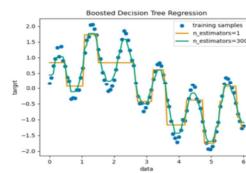


Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...



Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



Dans les Statistiques pour la SAé, on va voir une méthode de Régression linéaire à plusieurs variables, qui est une extension à plusieurs dimensions de la régression linéaire vue dans le Cours 3.



2. Comment ça, plusieurs dimensions ?

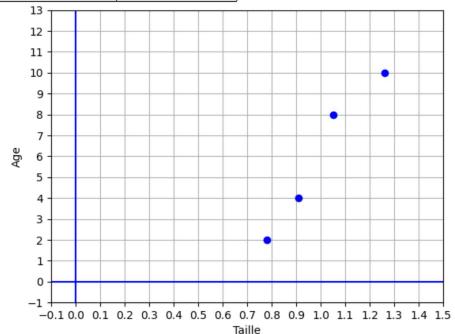
On travaille avec **plusieurs variables statistiques sur la même population** :

- 2 variables comme dans le Cours 3 sur les corrélations
- ... ou plus de 2.

On ne pourra donc pas forcément représenter le nuage de points !

Exemple 1 : Âge et taille des enfants (avec 2 variables statistiques)

	Enfant A	Enfant B	Enfant C	Enfant D	Enfant E
Taille	0.84	0.98	1.14	1.32	1.44
Âge	2	4	8	10	12



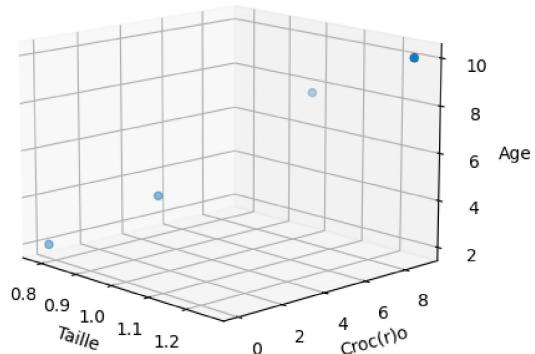
Exemple 1' : Âge et taille et prononciation des enfants (3 variables statistiques)

On ajoute une donnée pour chaque enfant : la qualité de sa prononciation du mot croc(r)odile :

	Enfant A	Enfant B	Enfant C	Enfant D
Taille	0.78	0.91	1.05	1.26
Croc(r)o	0	3	8	9
Âge	2	4	8	10

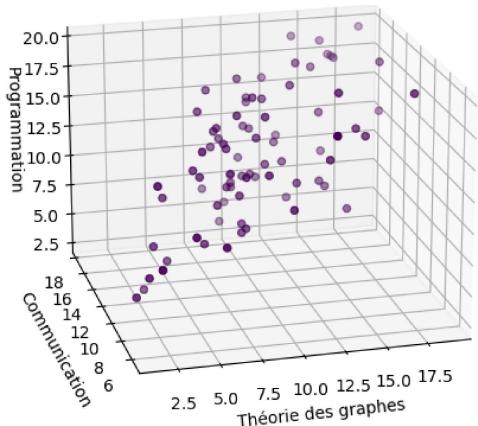
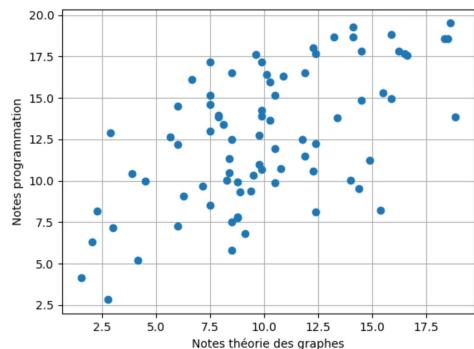
On a maintenant **3 variables statistiques sur la même population** :

- X : taille de l'enfant
- Y : prononciation de croc(r)odile
- Z : âge de l'enfant
- Population : enfants A,B,C et D



Exemple 2 : Notes S2 2017 (avec 2 ou 3 variables statistiques)

- Population : étudiant.es de 1e année de 2017
- X : notes en programmation orientée objets
- Y : notes en théorie des graphes
- Z : notes en communication



Si on ajoute les notes de Statistique, ou d'Anglais... on ne peut plus faire de représentation graphique !

code_nip	rg	moy	ue11	ue12
92553386554	29	12	4	16
59687716148	56	10	7	7
86157256760	16	13	36	14
37643613516	9	14	59	15
61148535285	75	11	22	9
82276273136	42	11	47	8
82318847777	31	12	21	10
18566782193	35	12	75	13
83264610326	9	14	52	11
22479873324	58	11	95	9
36235442780	22	13	69	14
76162577479	33	12	88	12
44371653331	76	11	19	8
69441313267	86	10	63	4
36565736237	43	12	45	8
36565736237	61	10	23	5
31521786814	36	11	74	11
18671253346	101	9	56	7
72254362779	18	13	79	15
86428578868	74	11	32	11
65167573403	20	12	98	9
63603868825	68	9	85	9
35634620712	26	12	56	10
47593183500	14	13	93	14
41716377246	17	13	32	18
81765279817	21	12	9	10
51783455648	24	12	62	16

Exemple 3 : Une vue extraite de la BD de SAé

- Population :
- Nombre de séries statistiques :
- Descripteurs :



3. Normalisation des données : centrer, réduire

La première étape de préparation de ce type de données sera de **normaliser** toutes les données, afin d'éviter que des échelles différentes ne donnent plus de poids à certaines variables qu'à d'autres.

Normalisation : centrer, réduire

Pour chaque variable statistique X , on calcule la moyenne \bar{X} et l'écart-type $\sigma(X)$ de la variable sur notre population, et pour chaque valeur $X(i)$:

- **on centre** en soustrayant la moyenne de la variable : $X(i) - \bar{X}$,
- **on réduit** en divisant le résultat par l'écart-type : $\frac{X(i) - \bar{X}}{\sigma(X)}$.

Exemple : normaliser 1 variable statistique

On prend une série de notes de 10 étudiant.es :

16, 11, 14, 10, 17, 12, 15, 9, 12, 14.

La moyenne sur ces 10 étudiant.es est 13, l'écart-type vaut 2.49.

- **on centre** : on soustraie la moyenne (13) à la note de chaque étudiant.e :

3, -2, 1, -3, 4, -1, 2, -4, -1, 1.

Par exemple pour le 1^e étudiant.e, on a remplacé 16 par $16 - 13 = 3$.

- **on réduit** : on divise ce qu'on a obtenu par l'écart-type 2.49 :

1.2, -0.8, 0.4, -1.2, 1.6, -0.4, 0.8, -1.6, -0.4, 0.4.

(résultats arrondi à 1 décimale)

Par exemple pour le 1^e étudiant.e, on a remplacé sa note centrée 3 par $3 \div 2.49 \approx 1.2$.

Exemple des âges, tailles et prononciation des enfants

	Enfant A	Enfant B	Enfant C	Enfant D
Taille	0.78	0.91	1.05	1.26
Croc(r)o	0	3	8	9
Âge	2	4	8	10

On calcule séparément la moyenne et l'écart-type pour chaque variable statistique :

- la taille X : $\bar{X} = 1$, $\sigma(X) \approx 0.18$
- la prononciation Y : $\bar{Y} = \dots$, $\sigma(Y) \approx 3.7$
- l'âge Z : $\bar{Z} = \dots$, $\sigma(Z) \approx 3.16$

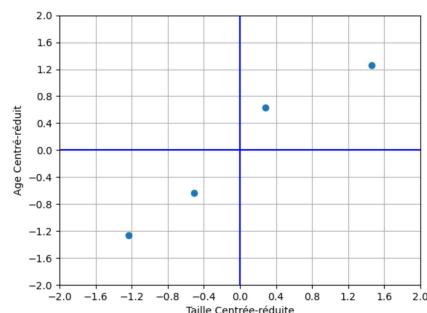
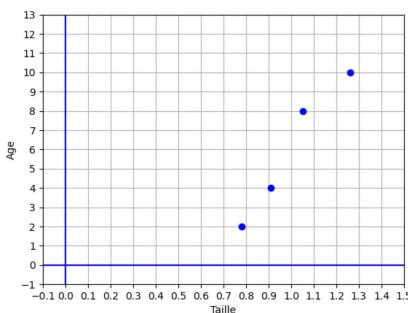
On centre-réduit chaque variable/ligne séparément :

	Enfant A	Enfant B	Enfant C	Enfant D
Taille CR	$\frac{0.78-1}{0.18} \approx -1.22$	$\frac{0.91-1}{0.18} \approx -0.5$	$\frac{1.05-1}{0.18} \approx 0.28$	$\frac{1.26-1}{0.18} \approx 1.44$
Croc(r)o CR	$\frac{0-\dots}{0.37} \approx \dots$			
Âge CR				

Normalisation vs nuage de points

Normaliser ne change pas la forme du nuage de points.

Mais normaliser évite que le nuage soit très étalé dans une direction et resserré sur une autre direction.



Utilité de la normalisation

Normaliser est utile surtout lorsqu'on a des données avec des échelles très différentes (par exemple dans les données sur les Sangliers vues en TP).

Années	Nb_sanglier_prelevés	Montant_indemnisations	Consommation_viande_porc	Nb_permis_chasse
2000	380518	21000000	35111	1425136
2001	401338	26000000	35564	1407874
2002	442466	24000000	36373	1394341
2003	475713	23500000	36691	1374183



4. La régression linéaire multiple

scikit-learn

Machine Learning in Python

[Getting Started](#)

[Release Highlights for 1.4](#)

[GitHub](#)

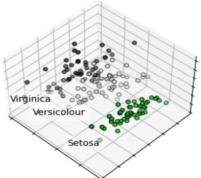
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: PCA, feature selection, non-negative matrix factorization, and more...

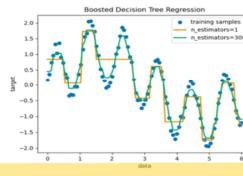


Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...



Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...



Régression

En mathématiques, la régression recouvre plusieurs méthodes permettant d'approcher une variable à partir d'autres qui lui sont corrélées. Par extension, le terme est aussi utilisé pour certaines méthodes d'ajustement de courbe.

Régression linéaire "simple" = avec 2 variables statistiques

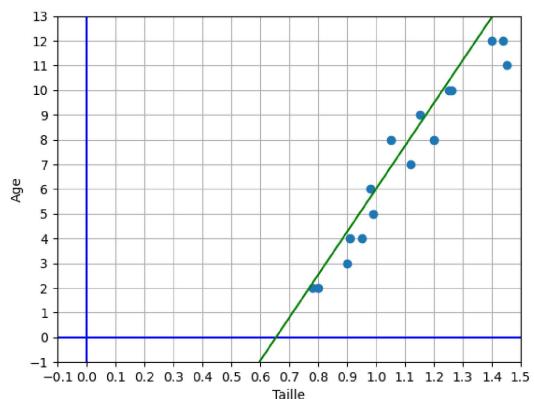
	Enfant A	Enfant B	Enfant C	Enfant D
Taille	0.78	0.91	1.05	1.26
Âge	2	4	8	10

La droite ci-contre a une équation de la forme $y = ax + b$, où a et b sont donnés par les formules du Cours 3 :

$$y \approx 17x - 11.$$

Ainsi, on peut trouver une approximation de l'âge d'un enfant à partir de sa taille :

$$\text{Age} \approx 17 \times \text{Taille} - 11$$



Régression linéaire avec 3 variables statistiques

Pour avoir plus de précision sur l'âge lorsqu'il y a plus d'enfants, on peut souhaiter rajouter une variable statistique. Par exemple, la qualité de la prononciation du mot *Croc(r)odile*, sur une échelle allant de 0 et 10 :

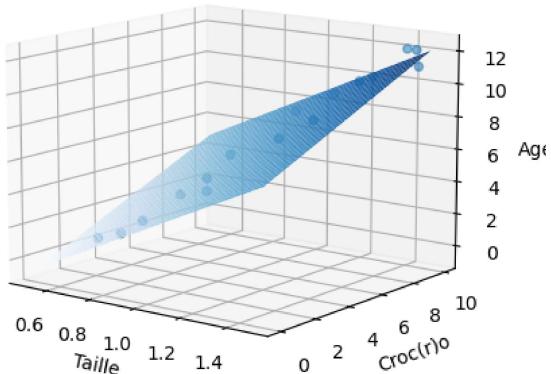
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Taille	0.84	0.98	1.14	1.32	1.44	0.8	0.95	1.2	1.25	1.4	0.9	0.99	0.98	1.12
Crocro	0	3	8	9	10	1	4	7	10	10	1	3.5	5	6
Âge	2	4	8	10	12	2	4	8	10	12	3	5	6	7

Dans un tel cas on peut trouver un plan, d'équation $y = a_1x_1 + a_2x_2 + b$ passant de manière optimale par le nuage de points.

Ici on trouvera : $y \approx 7x_1 + 0.5x_2 - 4$.

Ainsi, on peut trouver une approximation de l'âge d'un enfant à partir de sa taille et de sa prononciation :

$$\text{Age} \approx 7 \times \text{Taille} + 0.5 \times \text{Crocro} - 4$$



Régression linéaire avec 4 séries statistiques... ou plus ?

Pour avoir encore plus de précision sur l'âge des enfants, on peut encore rajouter des données. Par exemple, la capacité à se moucher (ou s'habiller) seul.e sur une échelle allant de 0 et 1 :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Taille	0.84	0.98	1.14	1.32	1.44	0.8	0.95	1.2	1.25	1.4	0.9	0.99	0.98	1.12
Crocro	0	3	8	9	10	1	4	7	10	10	1	3.5	5	6
Moucher	0	0.2	0.9	0.8	1	0.1	0.3	0.8	0.9	1	0.1	0.4	0.4	0.5
Âge	2	4	8	10	12	2	4	8	10	12	3	5	6	7

On ne peut alors plus faire de représentation graphique...

Mais on peut toujours trouver une relation linéaire !

Dans un tel cas on peut trouver une équation $y = a_1x_1 + a_2x_2 + a_3x_3 + b$ représentant de manière optimale la relation entre l'âge des enfants y , et les 3 autres variables x_1 , x_2 et x_3 (la taille, la prononciation de *Croc(r)odile*, et le mouchage).

Ici on trouvera : $y \approx 7x_1 + 3x_2 + x_3 - 4$.

Ainsi, on peut trouver l'âge approximatif des enfants par la formule :

$$\text{Age} \approx 7 \times \text{Taille} + 0.3 \times \text{Crocro} + 1 \times \text{Moucher} - 4$$

Régression linéaire multiple

- avec 2 variables X et Y , équation de droite $y = ax + b$
- avec 3 variables X_1, X_2 et Y , équation de plan $y = a_1x_1 + a_2x_2 + b$
- avec 4 variables X_1, X_2, X_3 et Y , équation $y = a_1x_1 + a_2x_2 + a_3x_3 + b$
- ...

Définition

On suppose qu'on a $n + 1$ variables statistiques : X_1, X_2, \dots, X_n et Y définies sur une même population.

Faire la **régression linéaire multiple** de Y en fonction de X_1, X_2, \dots, X_n consiste à trouver les coefficients a_1, a_2, \dots, a_n et b tels que l'équation

$$y = a_1x_1 + a_2x_2 + \dots + a_nx_n + b$$

soit la meilleure approximation possible de la relation entre Y et les variables X_1, X_2, \dots, X_n .

Définitions

Les variables statistiques X_1, X_2, \dots, X_n sont appelées les **variables explicatives**.

La variable statistique Y est appelée la **variable endogène**.

Les coefficients a_1, a_2, \dots, a_n et b de l'équation trouvée sont appelés les **paramètres**.

Exemple des âges, tailles et prononciations des enfants :

- on a bien des séries statistiques définies sur une même population :
.....
- la variable endogène est
- les variables explicatives sont
.....

En pratique en BUT1, on utilisera Python pour calculer les paramètres a_1, a_2, \dots, a_n de la régression linéaire multiple, on n'apprendra pas comment les calculer (sauf pour les + rapides en fin de TP2).



5. Analyse des paramètres obtenus

A partir de 4 variables statistiques, l'équation qu'on obtient est celle de ce qu'on appelle un *hyperplan*... qu'on ne peut pas représenter graphiquement !

Il faut donc savoir interpréter les **paramètres**.

Interprétation des paramètres

On suppose qu'on a fait la régression linéaire multiple d'une série Y en fonction de X_1, X_2, \dots, X_n , et qu'on a obtenu les paramètres a_1, a_2, \dots, a_n et b .

- si $a_i > 0$ est positif, cela signifie que la variable X_i influence positivement la variable Y .
- si $a_i < 0$, la variable X_i influence négativement Y .

De plus, si les variables X_1, X_2, \dots, X_n et Y ont été centrées-réduites avant d'effectuer la régression:

- $b = 0$
- plus $|a_i|$ est grand, plus la variable X_i a une influence forte sur Y .

Exemple des âges des enfants

- En faisant la régression linéaire multiple *sans centrer-réduire*, on obtient les coefficients $a_1 = 7.5, a_2 = 3.6, a_3 = 1.6$ et $b = -4$.

$$\text{Age} \approx 7.5 \times \text{Taille} + 0.4 \times \text{Crocro} + 1.6 \times \text{Moucher} - 4$$

Interprétation : les 3 variables influencent positivement l'âge.

C'est-à-dire que l'âge augmente quand la taille, la prononciation et la capacité de mouchage augmentent.

- En faisant la régression linéaire multiple après avoir centré-réduit, on obtient $a_1 = 0.47, a_2 = 0.37, a_3 = 0.17$ et $b = 0$.

Interprétation : c'est la taille qui influence le plus l'âge, et la capacité de mouchage qui influence le moins.



6. Coefficient de corrélation multiple

Pour les corrélations à 2 variables, on avait appris à calculer le coefficient de corrélation, qui indiquait si les données étaient proches de la droite.

Pour estimer si la régression linéaire multiple est de bonne qualité, on va calculer un coefficient de corrélation... multiple.

Coefficient de corrélation multiple

Soit $X_1, X_2 \dots, X_n$ et Y des variables statistiques définies sur une même population P de N individus.

On suppose qu'on a effectué la régression linéaire multiple de Y en fonction de X_1, X_2, \dots, X_n , et qu'on a obtenu les paramètres a_1, a_2, \dots, a_n et b .

- on introduit la variable statistique Y_{pred} qui, pour chaque individu i de la population P , donne la prédiction de la régression linéaire multiple, c'est-à-dire :

$$Y_{pred}(i) = a_1X_1(i) + a_2X_2(i) + \dots + a_nX_n(i) + b$$

- le **coefficient de corrélation multiple** de cette régression est alors

$$(Cor((X_1, \dots, X_n), Y))^2 = 1 - \frac{\sum_{i \in P} (Y_{pred}(i) - Y(i))^2}{N Var(Y)}$$

Remarque : Lorsqu'on n'a que 2 variables X et Y , ce calcul redonnera bien le coefficient $Cor(X, Y)$ vu dans le Cours 3.

Exemple des âges des enfants

On avait obtenu l'équation :

$$\text{Age} \approx 7.5 \times \text{Taille} + 0.4 \times \text{Crocro} + 1.6 \times \text{Moucher} - 4$$

Donc pour chaque enfant l'âge prédit est

$$Age_{pred} = 7.5 \times \text{Taille} + 0.4 \times \text{Crocro} + 1.6 \times \text{Moucher} - 4$$

Enfant	A	B	C	D	E	F	G	H	I	J	K	L	M	N
Âge	2	4	8	10	12	2	4	8	10	12	3	5	6	7
Âge prédit	1.5	3.9	7.9	9.7	11.7	2.2	4.7	8.5	10.1	11.4	3	5	5.5	7

On rappelle que la variance de l'âge est de 11.2. Donc

$$(Cor((\text{Taille}, \text{Crocro}, \text{Moucher}), \text{Age}))^2 = \dots \dots$$



Dernier QCM Moodle

Vous trouverez ce QCM sur l'espace Moodle de la ressource R2.08, section QCM.

A faire avant **mardi 3/06 23h59**.

Il sera composé des questions suivantes :

1. On vous donne 1 série statistique X sur une petite population (3 ou 4 valeurs), ainsi que l'écart-type de cette série. Vous devez donner la valeur normalisée pour un individu de cette série.
2. On vous donne un problème avec plusieurs variables statistiques. Vous devez indiquer laquelle est la variable endogène (à l'aide du problème formulé).