Les dangers de l'interprétation du coefficient de corrélation, le concept de corrélation partielle et le paradoxe de Simson

La corrélation observée entre deux variables X et Y peut être artificiellement masquée ou surévaluée en raison d'une ou plusieurs variables Z dans un système de données multidimensionnels.

La conséquence est possiblement une modification importante de la corrélation entre 2 variables Y et X quand on tient en compte une troisième variable Z. Dans ce cas l'interprétation des données peut être affectée et doit être corrigée. La variable Z est alors reconnue comme jouant un rôle de variable confondante.

Une modification importante dans un coefficient c'est un

- **changement de signe**
- **changement important de sa valeur : grande valeur vers petite valeur ou l'inverse**
- **obtenir un résultat près de zéro de la corrélation si on tient compte d'une troisième variable**

Selon la nature de/des variable(s) confondante(s), l'analyse statistique est différente.

La variable confondante Z est une échelle d'intervalle (quantitative)

Le principe est alors de calculer un coefficient de corrélation partielle en retirant la variance qui est due à une troisième variable Z (corrélation partielle entre X et Y notée alors rXY.Z). Cet indice de corrélation partielle permet par exemple de calculer la corrélation entre deux variables Y et X après avoir retiré l'effet d'une troisième variable Z, c'est à dire après avoir retiré l'influence de Z sur les variables Y et X due à Z. La formule de calcul est simple :

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2} \times \sqrt{1 - r_{yz}^2}} \qquad (1)$$

Lorsque qu'il existe plusieurs variables confondantes qui sont des échelles d'intervalles, la corrélation partielle est alors une corrélation partielle d'ordre p (rxy.z1z2...zp) et la formule est alors plus complexe. Il est souvent préférable si p est supérieur à 3 de passer par des techniques de régression.

Origine de la formule de corrélation partielle

Il est facile de montrer que le coefficient de corrélation ne change pas si les variables impliquées sont soumises à une transformation affine c.-à-d. que chaque variable est transformée selon la formule suivante. Par exemple Y est transformée en Y' selon la formule Y' = aY + b ou a et b sont des nombres réels. En particulier, si les variables impliquées sont centrées Y' = Y – Ybar avec a =1 et b= -Ybar, ou Ybar étant la moyenne de Y. Si on applique cette opération aux variables Y et X, elles deviennent de nouvelles variables disons

$$Y_z = Y - a\,Z \qquad (2)$$
$$X_z = X - b\,Z$$

ou a et b sont les coefficients de régression de Y sur Z et de X sur Z.

S signifie la sommation de toutes ls quantités sur l'ensemble de valeurs des variables impliquées.

Les coefficients a et b se calculent è l'aide des formules suivantes

$$a = S_{YZ} / S^2_Z \qquad (3)$$
$$b = S_{XZ} / S^2_Z$$

La corrélation partielle entre Y et X est la corrélation simple après avoir enlevé l'effet de la variable Z sur les variables originales Y et X c'est à dire la corrélation simple entre Y_z et X_z.
Cette corrélation a pour expression

$$S_{Y_z X_Z} / S_{YZ}\, S_{XZ} \qquad (4)$$

Les calculs des quantités sont obtenus par les formules suivantes :

$$S_{Y_z X_z} = S_{YZ} - a\,S_{YZ} - b\,S_{XZ} + ab\,S_Z^2$$

$$S_{Y_z}^2 = S_y - 2\,a\,S_{yz} + a^2 S_x \qquad (5)$$

$$S_{Y_x}^2 = S_y - 2\,b\,S_{yx} + b^2 S_x$$

Substituant les valeurs pour a et b et en simplifiant on arrive à la formule (1)

Exemple et implication en régression simple et multiple

Le tableau 1 représente les résultats de la note de huit élèves notés Y_note en fonction de trois variables potentiellement explicatives de la variabilité de Y_note c.-à-d. X1_poids, X2_age, X3_assiduité.

Première analyse : relation entre Y_note et assiduité

	1 élève	2 X1_poids	3 X2_age	4 X3_assiduité	5 Y_note
e1	e1	52	12,0	12	5,0
e2	e2	59	12,5	9	5,0
e3	e3	55	13,0	15	9,0
e4	e4	58	14,5	5	5,0
e5	e5	66	15,5	11	13,5
e6	e6	62	16,0	15	18,0
e7	e7	63	17,0	12	18,0
e8	e8	69	18,0	9	18,0

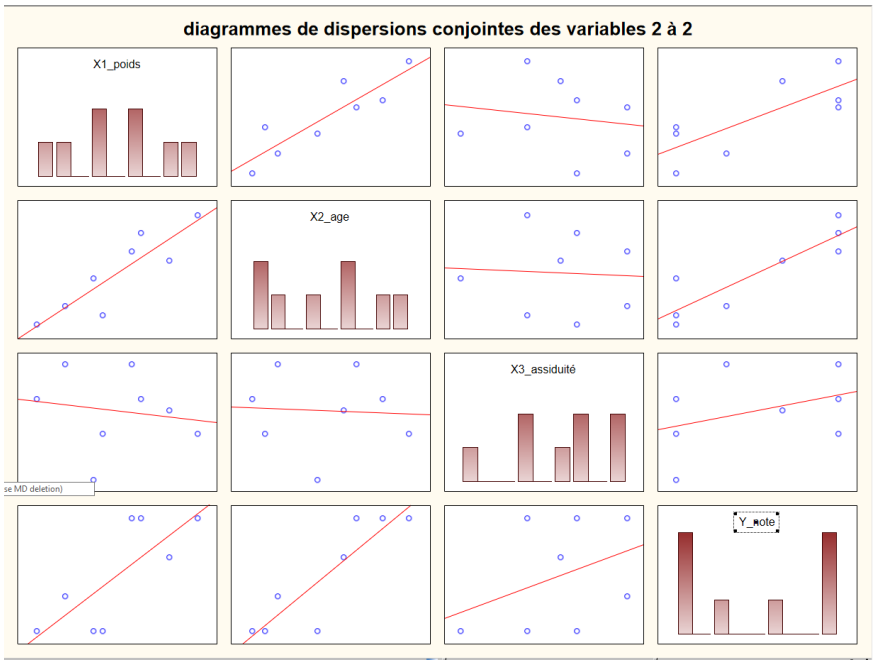
Tableau 1 : notes de huit élèves Y_note en fonction de trois variables X1_poids, X2_age, X3_asiduité

On s'interroge principalement de la relation et de l'influence de X3_assiduité sur la Y_note. On procède donc à une analyse de la corrélation entre les deux variables. Le graphique 1 représente la dispersion conjointe entre les variables dans lequel on a calculé le coefficient de corrélation entre les deux variables ainsi que l'équation de régression.

Le tableau 2 présente les corrélations entre les 4 variables X1_poids X2_age, X3_assiduité et Y_note

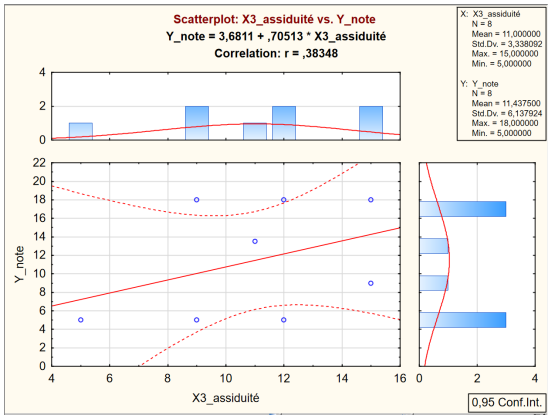
Variable	Correlations - variables dans unités originales					
	Means	Std.Dev.	X1_poids	X2_age	X3_assiduit	Y_note
X1_poids	60,50	5,63	1,00	0,88	-0,16	0,77
X2_age	14,81	2,19	0,88	1,00	-0,06	0,89
X3_assiduité	11,00	3,34	-0,16	-0,06	1,00	0,38
Y_note	11,44	6,14	0,77	0,89	0,38	1,00

Tableau 2 : corrélations entre les variables



Graphique 1 : diagramme de dispersion conjointe des variables

Le graphique 2 présente le graphique des variations de Y_note et X3_assiduité.
On a obtenu r(Y_note, X3_assiduité) = 0,38 une valeur qui nous apparait relativement faible au premier abord.



Graphique 2 : de dispersion conjointe et équation de régression entre Y_note et X3_assiduité

Deuxième analyse

On procède au calcul de la corrélation entre les deux variables Y_note et X3_assiduité mais cette fois en enlevant l'influence de la variable X2_age. Le concept de corrélation partielle est exactement ce qu'il faut utiliser. Le graphique 2 présente les résultats de la régression de Y_note en fonction de X3_assiduité mais cette fois on a enlevé l'effet de l'âge, ce qui nous parait logique dans les circonstances.

Les calculs est présenté dans le tableau 2. Nous avons transformé le tableau 1 initial dans le nouveau Tableau 2 ans lequel toutes les variables sont transformées sous forme centrées-réduites.

	1 élève	2 X1_poids	3 X2_age	4 X3_assiduité	5 Y_note	6	7 X1_poids_cr	8 X2_age_cr	9 X3_assiduité_cr	10 Y_note_cr
1	e1	52	12,0	12	5,0	données	-1,51	-1,29	0,30	-1,05
2	e2	59	12,5	9	5,0	transformées	-0,27	-1,06	-0,60	-1,05
3	e3	55	13,0	15	9,0	sous forme	-0,98	-0,83	1,20	-0,40
4	e4	58	14,5	5	5,0	centrée-réduite	-0,44	-0,14	-1,80	-1,05
5	e5	66	15,5	11	13,5	moyenne = 0	0,98	0,31	0,00	0,34
6	e6	62	16,0	15	18,0	écart-type =1	0,27	0,54	1,20	1,07
7	e7	63	17,0	12	18,0	écart-type =1	0,44	1,00	0,30	1,07
8	e8	69	18,0	9	18,0		1,51	1,46	-0,60	1,07

Tableau 2 : données des observations en variables originelles et variables centrées-réduites

Calcul des corrélations entre les variables centrées réduites (cr)

Variable	corrélations de variables centrées-réduites					
	Means	Std.Dev.	X1_poids_cr	X2_age_cr	X3_assiduité_cr	Y_note_cr
X1_poids_cr	0,00	1,00	1,00	0,88	-0,16	0,77
X2_age_cr	0,00	1,00	0,88	1,00	-0,06	0,89
X3_assiduité_cr	0,00	1,00	-0,16	-0,06	1,00	0,38
Y_note_cr	0,00	1,00	0,77	0,89	0,38	1,00

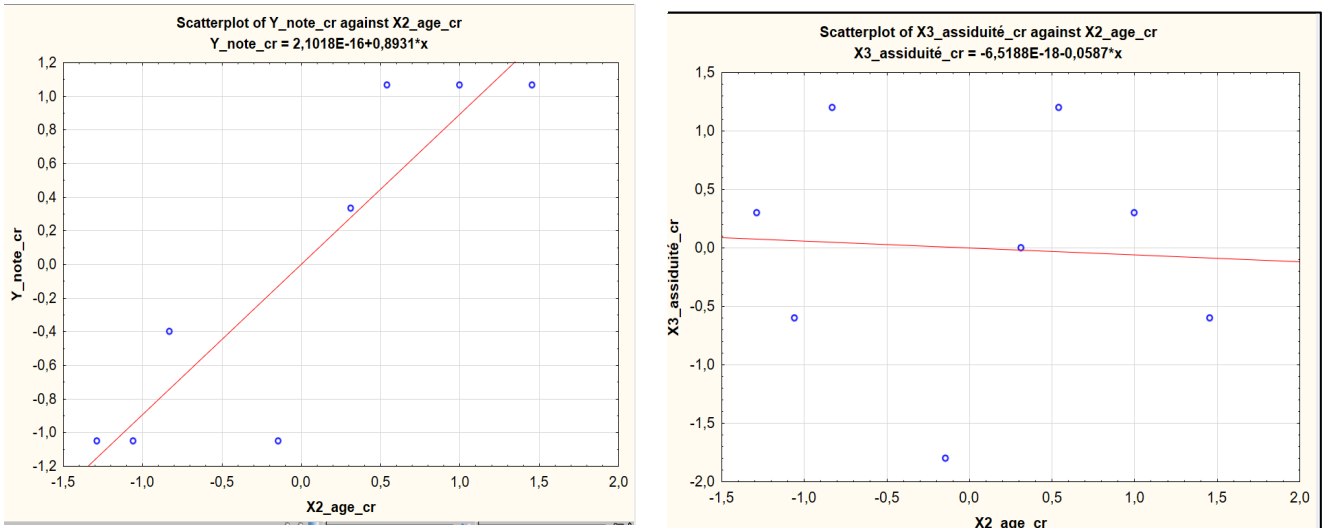
Tableau 3 : corrélations

Les valeurs des corrélations du tableau 3 sont identiques à celles du tableau 2.

Nous allons maintenant enlever l'effet de la variable X2_âge sur la variable Y_note et la variable X3_assiduité en régressant chacune d'entre elles sur X2_age. Nous obtenons les équations de régression :

$$Y\_note\_cr = 0{,}891 * X2\_age\_cr \quad (6)$$

$$X3\_assiduité\text{-}cr = 0{,}069 * X2\_age\_cr$$



Graphique 3 : graphiques de dispersion conjointe et régressions de Y_note_cr et X3_assiduité sur X3_age_cr

Le résultat de cette opération est donné dans le tableau 4.

8 X2_age_cr	9 X3_assiduité_cr	10 Y_note_cr	11	12 X3_cr_corrigée_age	13 Y_note_cr_corrigée_age
-1,29	0,30	-1,05	variables	0,39	0,39
-1,06	-0,60	-1,05	corrigées	-0,53	-0,53
-0,83	1,20	-0,40	pour effet	1,26	1,26
-0,14	-1,80	-1,05	de l'âge	-1,79	-1,79
0,31	0,00	0,34		-0,02	-0,02
0,54	1,20	1,07		1,16	1,16
1,00	0,30	1,07		0,23	0,23
1,46	-0,60	1,07		-0,70	-0,70

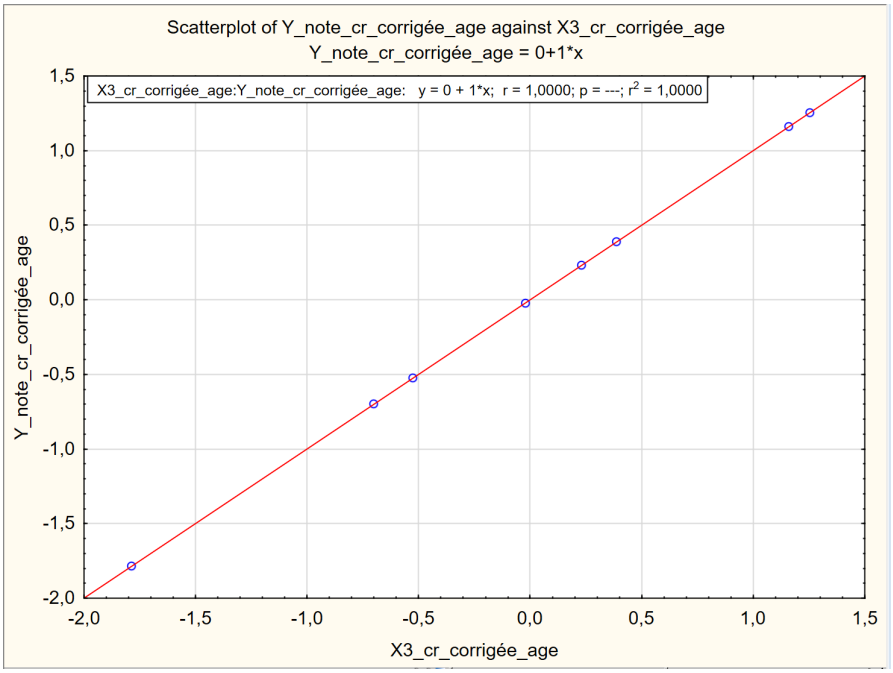
Tableau 4 : valeurs corrigées pour l'effet de l'âge

Nous sommes en présence de nouvelles variables de X_note_age et X3_assiduité_age car elles furent corrigées pour l'effet de l'âge. Le calcul de la corrélation entre les variables corrigées après avoir enlevé l'effet de l'âge donne 1. La corrélation entre la variable Y_note et X3_asiduité a augmenté à 1 de 0,38 qu'elle était quand on ne tenait pas en compte de la troisième variable X2_age. Cette nouvelle valeur fait beaucoup plus de sens que la valeur de 0,38 quand on faisait abstraction de l'âge dans le calcul.

La variable X2_age peut recevoir le qualificatif -de confondante car la corrélation entre X_note et X3_assiduité change considérablement quand on tient en compte une troisième

variable X2_age. Ce phénomène porte le nom de paradoxe de Simpson. Il faut être conscient de sa présence potentielle dans l'analyse de données multidimensionnelles.

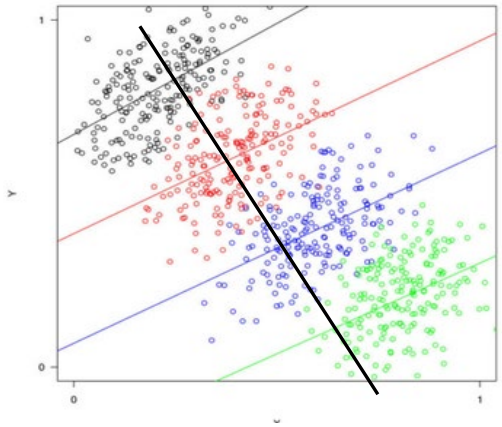
Y_note_cr_corrigée_age: y = 0 + 1*X3_cr_corrigée_age; r = 1,0000; -; r^2 = 1,00 **(8)**



<u>Graphique</u> 4 : corrélation entre Note et Assiduité corrigée par l'âge

La variable confondante Z est qualitative (catégorique)

Une variable qualitative ((qui permet de distinguer différents groupes) conduit à calculer la corrélation pour chaque groupe. On peut ainsi avoir des surprises avec par exemple une corrélation négative entre deux variables x et y, qui devient positive pour chacun des groupes (c'est une des expressions du paradoxe de Simpson*, cf. graphique-5) ou encore des corrélations qui varient selon les groupes et qui sont très différentes de celles observées globalement. La droite globale est à une corrélation négative sur l'ensemble des données mais si on fait l'analyse par sous ensemble les corrélations observées sont positives.



Graphique-5 : Illustration du paradoxe de Simpson

La corrélation entre X et Y est négative (nuage de points orienté vers la gauche) mais pour les 4 groupes distingués par la variable Z (4 groupes d'âge correspondant aux 4 couleurs dans le nuage **de points), les corrélations entre X et Y sont toutes positives !**

Paradoxe de Simpson

Le paradoxe de Simpson (ou effet de Yule-Simpson) a été décrit initialement par Udny Yule en 1903 puis repris par Edward Simpson en 1951. De façon générale, cet effet correspond à l'inversion d'un effet (fréquence de guérison, corrélation, etc.) observé dans plusieurs groupes lorsque l'on regroupe toutes les données (par exemple une différence de moyennes entre deux conditions est positive dans un premier groupe, positive dans le second groupe mais s'inverse quand on combine les deux groupes).

Il est très important de prendre en compte une variable confondante surtout lorsque l'on a des données provenant de différents groupes bien identifiés. En effet, on observe parfois des résultats très surprenants sur les moyennes, les fréquences (exemple les plus fréquents pour illustrer le paradoxe Simpson). Cependant, cet effet existe aussi pour les corrélations, la figure B-5 en est l'illustration. Ce paradoxe n'est pas vraiment un. Dans l'exemple donné (figure B-5) les moyennes des scores sur la variable X augmentent avec l'âge et alors qu'ils diminuent pour Y avec l'âge. Si l'on regarde la relation entre x et y tout âge confondu, la corrélation devient négative (alors qu'elle était positive pour chaque groupe d'âge). Cet effet particulier est donc toujours à prendre en compte surtout dans les études développementales.

Un vidéo avec des exemples pratiques du paradoxe de Simpson
https://www.youtube.com/watch?time_continue=11&v=vs_Zzf_vL2I

Le vidéo fait bien la distinction entre des études statistiques rétrospectives basées sur des données existantes déjà observées et stockées versus des études prospectives dont les données sont inexistantes et à recueillir comme dans études expérimentales. En d'autres termes, les dangers et les pièges potentiels de l'interprétation des données provenant du BIG DATA.