Ricco Rakotomalala

Analyse de corrélation

Étude des dépendances - Variables quantitatives

Version 1.0

Université Lumière Lyon 2

Avant-propos

Ce support décrit les méthodes statistiques destinées à quantifier et tester la liaison entre 2 variables quantitatives : on parle d'analyse de corrélation dans la littérature.

Il correspond à une partie des enseignements d'économétrie (je préfère l'appellation *Régression Linéaire Multiple*) en L3-IDS de la Faculté de Sciences Economiques de l'Université Lyon 2 (`http://dis.univ-lyon2.fr/`). Il se veut avant tout opérationnel. Nous nous concentrons sur les principales formules et leur mise en oeuvre pratique avec un tableur. Autant que possible nous ferons le parallèle avec les résultats fournis par les logiciels de statistique libres et/ou commerciaux. Le bien-fondé des tests, la pertinence des hypothèses à opposer sont peu ou prou discutées. Nous invitons le lecteur désireux d'approfondir les bases théoriques à consulter les ouvrages énumérés dans la bibliographie.

Un document ne vient jamais du néant. Pour élaborer mes supports, je m'appuie sur différentes références, des ouvrages disais-je plus tôt, mais aussi des ressources en ligne qui sont de plus en plus présents aujourd'hui dans la diffusion de la connaissance. Les seuls bémols par rapport à ces documents sont le doute que l'on pourrait émettre sur l'exactitude des informations prodiguées, mais la plupart de leurs auteurs sont des enseignants-chercheurs qui font sérieusement leur travail (de toute manière je multiple les vérifications avant d'y faire référence) ; une disponibilité plus ou moins aléatoire, au gré des migrations des serveurs et de la volonté de leurs auteurs, auquel il est très difficile de remédier (désolé s'il y a des liens qui ne fonctionnent plus) ; les informations sont disparates, avec une absence d'organisation, à la différence des ouvrages qui suivent une ligne pédagogique très structurante.

Néanmoins, ces ressources en ligne renouvellent profondément le panorama des documents disponibles pour les enseignements. Il y a la gratuité bien sûr. C'est un aspect important. Mais il y a aussi l'accès à des fonctionnalités qui sont moins évidentes avec les supports classiques. Par exemple, dans la grande majorité des cas, les données qui illustrent les documents sont accessibles sur le site web de diffusion. C'est un atout fort. Pour notre cas, le lecteur pourra (j'espère) reproduire aisément les calculs présentés à l'aide du fichier EXCEL qui accompagne ce document.

Concernant ce support, rendons à César ce qui lui appartient. Parmi les différentes références utilisées, j'ai beaucoup été influencé par 2 excellents ouvrages : celui de Chen et Popovitch [2], il fait partie de la non moins excellente série "Quantitative Applications in the Social Sciences" de Sage University Paper ;

celui de Aïvazian [1], qui fait partie des références, introuvables aujourd'hui, que je bichonne dans ma bibliothèque.

Ce support est totalement gratuit. Vous pouvez en reprendre des parties dans vos propres productions ou dans vos enseignements, tant qu'elles sont elles-mêmes diffusées à titre non commercial. Une citation de la source originale serait appréciée.

Bien entendu, selon la formule consacrée, ce document n'engage que son auteur. Toutes suggestions ou commentaires qui peuvent en améliorer le contenu sont le bienvenu.

Table des matières

Partie I

Analyse de Corrélation

1

Étudier la liaison entre deux variables quantitatives

1.1 Objectif : analyser la liaison

Soient X et Y deux grandeurs statistiques quantitatives observées. On souhaite :

1. Déterminer s'il existe une relation entre X et Y.
2. Caractériser la forme de la liaison (la relation) entre X et Y (positive ou négative, linéaire ou non linéaire, monotone ou non monotone).
3. Tester si la liaison est statistiquement significative.
4. Quantifier l'intensité de la liaison.
5. Valider la liaison identifiée. Est-ce qu'elle n'est pas le fruit d'un simple artefact ou le produit d'autres informations sous-jacentes dans les données ?

Attention, la position des variables est symétrique dans ce cadre. On ne veut pas évaluer l'influence d'une des variables sur l'autre, à la différence de la régression.

1.2 Analyse graphique

L'analyse graphique est une bonne manière de comprendre les différentes caractéristiques énumérées ci-dessus. Le graphique "nuage de points" est l'outil privilégié [1]. Nous plaçons en abcisse la variable X, en ordonnée la variable Y, chaque observation est positionnée dans le repère ainsi constitué. L'intérêt est multiple : nous pouvons situer les proximités entre les individus ; étudier la forme globale des points, voir notamment s'il existe une forme de liaison ou de régularité ; détecter visuellement les points qui s'écartent des autres, les observations atypiques ; vérifier s'il n'y a pas de regroupement suspects, laissant entendre qu'il y a en réalité une troisième variable qui influence le positionnement des individus...

Dans la figure 1.1, nous illustrons quelques types de liaisons qui peuvent exister entre 2 variables continues :

– *Liaison linéaire positive.* X et Y évoluent dans le même sens, une augmentation de X entraîne une augmentation de Y, du même ordre quelle que soit la valeur de X.

1. `http://www.ebsi.umontreal.ca/jetrouve/illustre/nuage.htm`

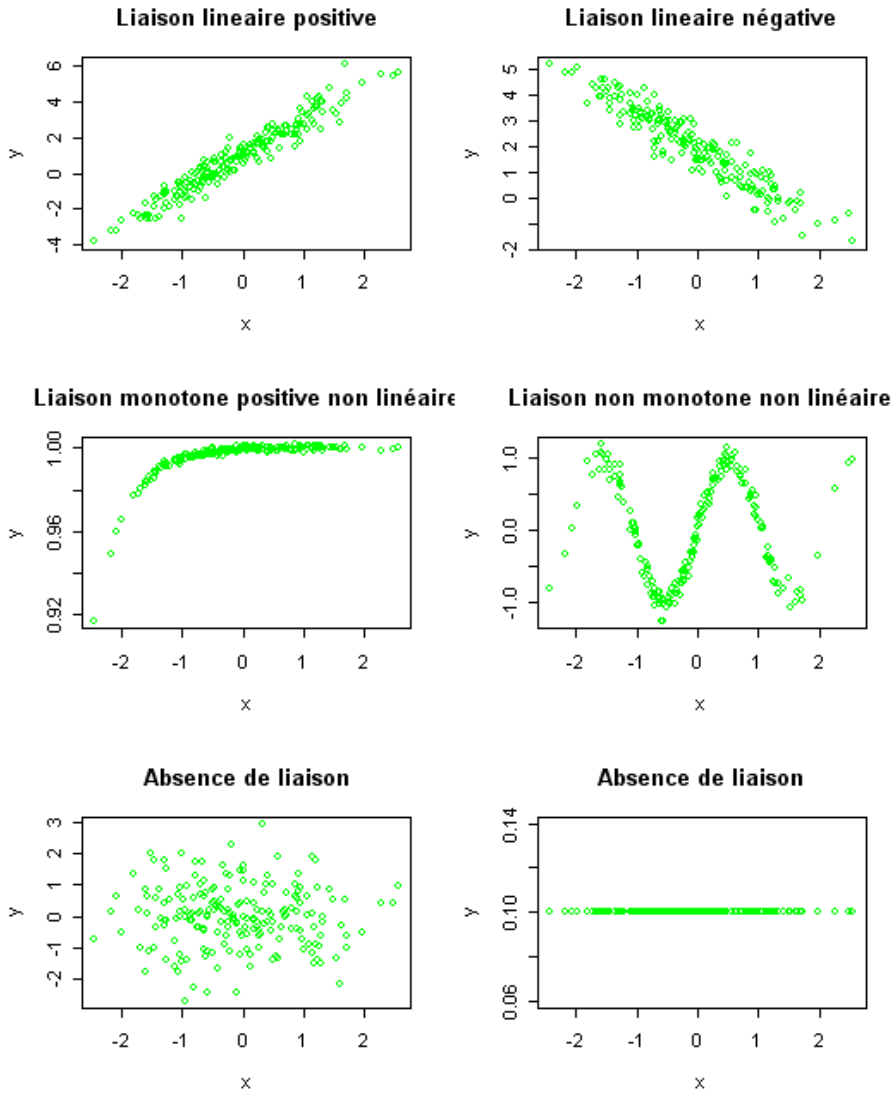


Fig. 1.1. Quelques types de liaisons entre 2 variables

- *Liaison linéaire négative.* X et Y évoluent en sens inverse. La pente est inchangée quelle que soit la valeur de X.
- *Liaison monotone positive non-linéaire.* X et Y évoluent dans le même sens, mais la pente est différente selon le niveau de X.
- *Liaison non-linéaire non-monotone.* Il y a une relation fonctionnelle (de type sinusoïdale ici) entre X et Y. Mais la relation n'est pas monotone, Y peut augmenter ou diminuer selon la valeur de X.
- *Absence de liaison.* La valeur de X ne donne indication sur la valeur de Y, et inversement. L'autre situation caractéristique est que X (ou Y) est constant quelle que soit la valeur de la seconde variable.

1.3 Notations

Nous utiliserons les conventions suivantes dans ce support :

- Une **variable** est notée en majuscules (X est une variable).
- x_i correspond à la **valeur** prise par l'observation numéro i pour la variable X.
- La population parente est notée Ω^{pop}.
- L'échantillon est noté Ω, l'effectif de l'échantillon est $n = card(\Omega)$. Dans le cadre de la corrélation, nous travaillons sur un échantillon de n observations, constituées de couples (x_i, y_i) c.-à-d. $\Omega = \{(x_i, y_i), i = 1, \ldots, n\}$.
- La moyenne empirique calculée sur l'échantillon est $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- L'écart type empirique est $s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

2

Coefficient de corrélation de Bravais-Pearson

2.1 Covariance

L'objectif de la covariance est de quantifier la liaison entre deux variables X et Y, de manière à mettre en évidence le *sens* de la liaison et son *intensité*.

2.1.1 Définition

La covariance est égale à l'espérance du produit des variables centrées.

$$COV(X,Y) = E\{[X - E(X)][Y - E(Y)]\} \tag{2.1}$$

On peut aussi l'écrire comme l'espérance du produit des variables, moins le produit des espérances.

$$COV(X,Y) = E[XY] - E[X]E[Y] \tag{2.2}$$

Signification. La covariance mesure la tendance des deux variables à être simultanément au dessus ou en dessous de leurs espérances respectives. Elle modélise une liaison monotone.

Quelques remarques :

1. La référence est donc l'espérance mathématique, on veut savoir si : lorsque X est supérieur a son espérance, Y a tendance à être supérieur (ou inférieur) à son espérance.
2. On peut maintenant quantifier le *sens* de la liaison
   - $COV(X,Y) > 0$: la relation est positive c.-à-d. lorsque X est plus grand que son espérance, Y a tendance à l'être également ;
   - $COV(X,Y) = 0$: absence de relation monotone ;
   - $COV(X,Y) < 0$: la liaison est négative c.-à-d. lorsque X est plus grand que son espérance, Y a tendance à être plus petit que sa propre espérance.
3. La covariance d'une variable avec elle-même est la variance, la relation est toujours positive. En effet,

$$\begin{aligned} COV(X,X) &= E\{[X-E(X)][X-E(X)]\} \\ &= E\{[X-E(X)]^2\} \\ &= V(X) \\ &> 0 \end{aligned}$$

2.1.2 Propriétés

Voici les principales propriétés de la covariance (Note : essayez d'effectuer les démonstrations à partir de la définition et des propriétés de l'espérance mathématique).

1. **Symétrie.** $COV(X,Y) = COV(Y,X)$
2. **Distributivité.** $COV(X,Y+Z) = COV(X,Y) + COV(X,Z)$ (... ne pas oublier que $E[X+Y] = E[X] + E[Y]$)
3. **Covariance avec une constante.** $COV(X,a) = 0$
4. **Covariance avec une variable transformée.**(Transformation affine) $COV(X, a + b \times Y) = b \times COV(X,Y)$
5. **Variance de la somme de deux variables aléatoires.** $V(X+Y) = V(X)+V(Y)+2\times COV(X,Y)$
6. **Covariance de 2 variables indépendantes.**

$$X, Y \text{independants} \Rightarrow COV(X,Y) = 0$$

Attention, la réciproque est généralement fausse. Ce n'est pas parce que la covariance est nulle que les variables sont forcément indépendantes.
(Remarque : Pour démontrer cette propriété, il ne faut pas oublier que lorsque X et Y sont indépendants, $E[X \times Y] = E[X] \times E[Y]$).

2.1.3 Domaine de définition

La covariance est définie dans l'ensemble des réels c.-à-d. $-\infty < COV(.) < +\infty$. Il permet de se rendre compte du sens de la liaison. Plus sa valeur est élevée (en valeur absolue), plus la liaison est forte. Mais nous ne savons pas quelle est la limite. Nous ne pouvons pas non plus comparer la covariance d'une variable X avec deux autres variables Y et Z. Dans la pratique, nous préférerons donc une mesure normalisée : le coefficient de corrélation répond à ces spécifications (section 2.2).

2.1.4 Estimation

Sur un échantillon de taille n, la covariance empirique est définie de la manière suivante :

$$\hat{S}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \tag{2.3}$$

On montre que c'est un estimateur biaisé de la covariance, en effet $E[\hat{S}_{xy}] = \frac{n-1}{n} COV(X,Y)$.

L'estimateur sans biais de la covariance [1] s'écrit par conséquent :

$$C\hat{O}V(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1} \tag{2.4}$$

Détails des calculs sur un exemple. Pour préciser les idées, détaillons les calculs dans le tableur EXCEL. Nous cherchons à calculer la covariance entre la cylindrée et la puissance de 28 véhicules (Figure 2.1) :

	A	B	C	D	E	F	G	H
1	Numero	Modele	Cylindree	Puissance	XY			
2	1	Daihatsu Cuore	846	32	27072			
3	2	Suzuki Swift 1.0 GLS	993	39	38727			
4	3	Fiat Panda Mambo L	899	29	26071			
5	4	VW Polo 1.4 60	1390	44	61160			
6	5	Opel Corsa 1.2i Eco	1195	33	39435			
7	6	Subaru Vivio 4WD	658	32	21056			
8	7	Toyota Corolla	1331	55	73205		Cov.Empirique	18381.4133
9	8	Opel Astra 1.6i 16V	1597	74	118178		Cov.Non-Biaisé	19062.2063
10	9	Peugeot 306 XS 108	1761	74	130314			
11	10	Renault Safrane 2.2. V	2165	101	218665			
12	11	Seat Ibiza 2.0 GTI	1983	85	168555		Cov.Excel	18381.4133
13	12	VW Golt 2.0 GTI	1984	85	168640			
14	13	Citroen ZX Volcane	1998	89	177822			
15	14	Fiat Tempra 1.6 Liberty	1580	65	102700			
16	15	Fort Escort 1.4i PT	1390	54	75060			
17	16	Honda Civic Joker 1.4	1396	66	92136			
18	17	Volvo 850 2.5	2435	106	258110			
19	18	Ford Fiesta 1.2 Zetec	1242	55	68310			
20	19	Hyundai Sonata 3000	2972	107	318004			
21	20	Lancia K 3.0 LS	2958	150	443700			
22	21	Mazda Hachtback V	2497	122	304634			
23	22	Mitsubishi Galant	1998	66	131868			
24	23	Opel Omega 2.5i V6	2496	125	312000			
25	24	Peugeot 806 2.0	1998	89	177822			
26	25	Nissan Primera 2.0	1997	92	183724			
27	26	Seat Alhambra 2.0	1984	85	168640			
28	27	Toyota Previa salon	2438	97	236486			
29	28	Volvo 960 Kombi aut	2473	125	309125			
30	n		Moyenne		Somme			
31	28		1809.07	77.71	4451219			

Fig. 2.1. Détails des calculs - Estimation de la covariance

- Au bas de la feuille de calcul, en colonne C et D nous avons la moyenne de chaque variable.
- Dans la colonne E, nous calculons le produit $(x_i y_i)$, dont la somme est 4451219.
- Nous pouvons alors former la covariance empirique (formule 2.3), elle est égale à 18381.4133.
- L'estimateur sans biais (formule 2.4) étant lui égal à 19062.2063. L'écart entre les deux valeurs s'amenuise à mesure que l'effectif n augmente.
- Notons que la fonction "COVARIANCE(...)" du tableur EXCEL fournit la covariance empirique.

Comparaison de covariances. Illustrons maintenant l'impossibilité de comparer des covariances lorsque les variables sont exprimées dans des unités différentes. Nous souhaitons travailler sur un fichier de 28 véhicules décrites à l'aide de la cylindrée, la puissance, le poids et la consommation (Figure 2.2 ; ce fichier reviendra plusieurs fois dans ce support).

1. Faire le parallèle avec l'estimateur sans biais de la variance

	A	B	C	D	E	F
1	Numero	Modele	Cylindree	Puissance	Poids	Conso
2	1	Daihatsu Cuore	846	32	650	5.7
3	2	Suzuki Swift 1.0 GLS	993	39	790	5.8
4	3	Fiat Panda Mambo L	899	29	730	6.1
5	4	VW Polo 1.4 60	1390	44	955	6.5
6	5	Opel Corsa 1.2i Eco	1195	33	895	6.8
7	6	Subaru Vivio 4WD	658	32	740	6.8
8	7	Toyota Corolla	1331	55	1010	7.1
9	8	Opel Astra 1.6i 16V	1597	74	1080	7.4
10	9	Peugeot 306 XS 108	1761	74	1100	9.0
11	10	Renault Safrane 2.2. V	2165	101	1500	11.7
12	11	Seat Ibiza 2.0 GTI	1983	85	1075	9.5
13	12	VW Golt 2.0 GTI	1984	85	1155	9.5
14	13	Citroen ZX Volcane	1998	89	1140	8.8
15	14	Fiat Tempra 1.6 Liberty	1580	65	1080	9.3
16	15	Fort Escort 1.4i PT	1390	54	1110	8.6
17	16	Honda Civic Joker 1.4	1396	66	1140	7.7
18	17	Volvo 850 2.5	2435	106	1370	10.8
19	18	Ford Fiesta 1.2 Zetec	1242	55	940	6.6
20	19	Hyundai Sonata 3000	2972	107	1400	11.7
21	20	Lancia K 3.0 LS	2958	150	1550	11.9
22	21	Mazda Hachtback V	2497	122	1330	10.8
23	22	Mitsubishi Galant	1998	66	1300	7.6
24	23	Opel Omega 2.5i V6	2496	125	1670	11.3
25	24	Peugeot 806 2.0	1998	89	1560	10.8
26	25	Nissan Primera 2.0	1997	92	1240	9.2
27	26	Seat Alhambra 2.0	1984	85	1635	11.6
28	27	Toyota Previa salon	2438	97	1800	12.8
29	28	Volvo 960 Kombi aut	2473	125	1570	12.7

Fig. 2.2. Fichier "consommation des automobiles"

La covariance empirique de la variable "consommation" avec les autres variables nous donne respectivement : cylindrée = 1197.6 ; puissance = 61.7 ; poids = 616.3. Manifestement, les valeurs ne se situent pas sur la même échelle, toute comparaison n'a aucun sens.

2.2 Coefficient de corrélation de Pearson

2.2.1 Définition

Le coefficient de corrélation linéaire simple, dit de *Bravais-Pearson* (ou de *Pearson*), est une normalisation de la covariance par le produit des écarts-type des variables.

$$r_{xy} = \frac{COV(X,Y)}{\sqrt{V(X) \times V(Y)}} \tag{2.5}$$

$$= \frac{COV(X,Y)}{\sigma_x \times \sigma_y} \tag{2.6}$$

Remarque 1 (Précisions sur la notation). Dans ce qui suit, s'il n'y a pas d'ambiguïtés, nous omettrons les indices X et Y.

2.2.2 Propriétés

1. Il est de même signe que la covariance, avec les mêmes interprétations.

2. X et Y sont indépendants, alors $r = 0$. La réciproque est fausse, sauf cas particulier que nous précisons maintenant.
3. Lorsque le couple de variables (X, Y) suit une loi normale bi-variée, et uniquement dans ce cas, nous avons l'équivalence $r = 0 \Leftrightarrow X$ et Y sont indépendants. Dans ce cas, le coefficient de corrélation caractérise parfaitement la liaison entre X et Y. Dans les autres cas, le coefficient de corrélation constitue une mesure parmi les autres de l'intensité de la corrélation.
4. Le coefficient de corrélation constitue une mesure de l'**intensité de liaison linéaire** entre 2 variables. Il peut être égal à zéro alors qu'il existe une liaison fonctionnelle entre les variables. C'est le cas lorsque la liaison est non monotone.
5. La corrélation d'une variable avec elle même est $r_{xx} = 1$.

2.2.3 Domaine de définition

Le coefficient de corrélation est indépendant des unités de mesure des variables, ce qui autorise les comparaisons. La mesure est normalisée, elle est définie entre [2]

$$-1 \leq r \leq +1 \tag{2.7}$$

Lorsque :
- $r = +1$, la liaison entre X et Y est linéaire, positive et parfaite c.-à-d. la connaissance de X nous fournit la valeur de Y (et inversement).
- $r = -1$, la liaison est linéaire et négative.

2.2.4 Quelques exemples graphiques

Reprenons les exemples graphiques présentés ci-dessus (section 1.2, figure 1.1), affichons maintenant le coefficient de corrélation (Figure 2.3). Si la liaison est non monotone, r n'est d'aucune utilité. Si la liaison est monotone mais non linéaire, r caractérise mal l'intensité de la liaison.

2.3 Coefficient de corrélation empirique

2.3.1 Définition

Sur un échantillon de taille n, nous estimons le coefficient de corrélation à l'aide de la formule suivante (Équation 2.8) :

2. Pour réaliser la démonstration, il faut s'appuyer sur deux pistes

$$V(\frac{X}{\sigma_x} + \frac{Y}{\sigma_y}) \geq 0 \Rightarrow r \geq -1$$
$$V(\frac{X}{\sigma_x} - \frac{Y}{\sigma_y}) \geq 0 \Rightarrow r \leq +1$$

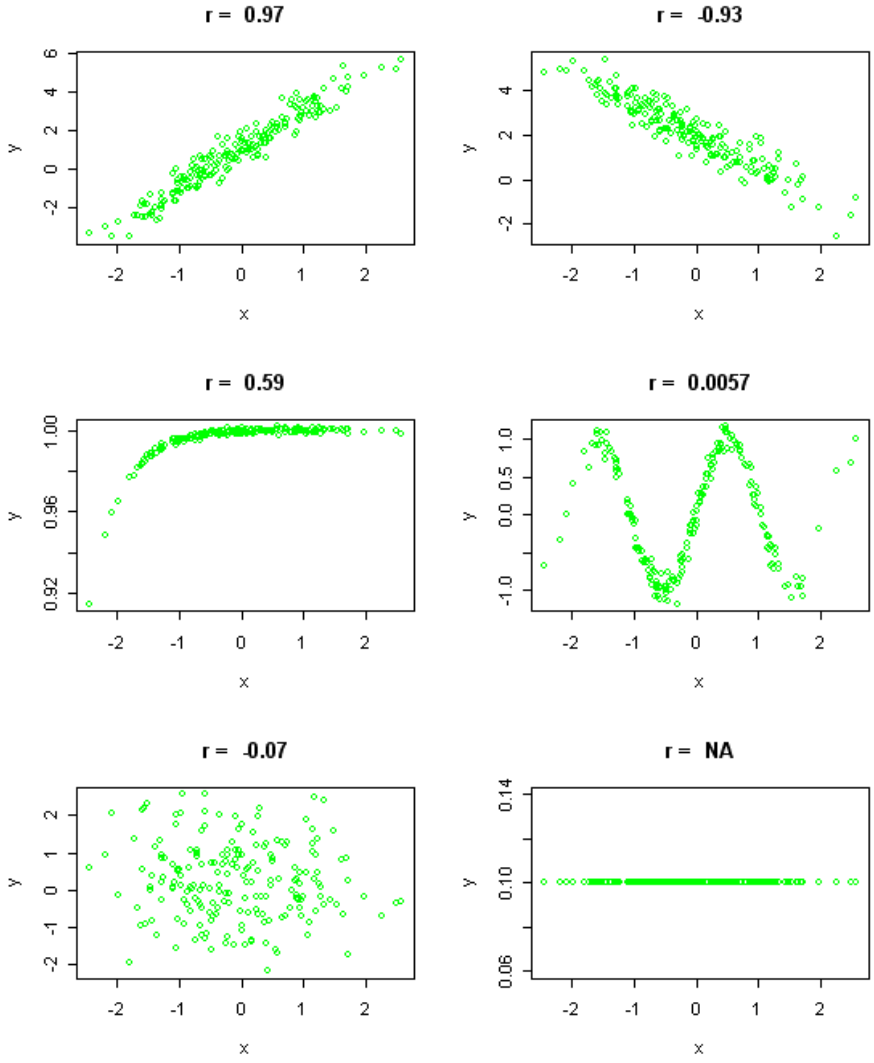


Fig. 2.3. Coefficients de corrélation pour différents types de liaison

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.8}$$

On parle de *coefficient de corrélation empirique* dans la littérature. Après quelques simplifications, nous pouvons également utiliser la formulation suivante :

$$\hat{r} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \times \sqrt{\sum y_i^2 - n\bar{y}^2}} \tag{2.9}$$

Nous pouvons calculer le coefficient de corrélation sans disposer du détail des observations, les quantités pré-calculées $\bar{x}$, $\bar{y}$, $\sum x_i y_i$, $\sum x_i^2$ et $\sum y_i^2$ suffisent.

2.3.2 Interprétation

Le coefficient de corrélation sert avant tout à caractériser une relation linéaire positive ou négative. Il s'agit d'une mesure symétrique. Plus il est proche de 1 (en valeur absolue), plus la relation est forte. $r = 0$ indique l'absence de corrélation, il équivaut à un test d'indépendance si et seulement si le couple (X, Y) suit une loi normale bivariée.

La valeur de $\hat{r}$ n'a pas de signification intrinsèque. En revanche, son carré c.-à-d. $\hat{r}^2$, que l'on appelle **coefficient de détermination**, s'interprète comme la **proportion de variance** de Y (resp. X) **linéairement expliquée** par X (resp. Y). On peut faire le rapprochement avec les résultats produits avec la régression linéaire [3].

Ainsi, $\hat{r} = 0.9$, on voit que la liaison est forte, puisqu'elle se rapproche de 1. C'est tout. En revanche, avec $\hat{r}^2 = 0.81$, on peut dire que 81% de la variance de Y est expliquée par X (et inversement)(voir [3], page 90).

Il existe par ailleurs d'autres interprétations du coefficient de corrélation de Pearson. Parmi les plus intéressants figure l'interprétation géométrique qui assimile r au cosinus de l'angle entre les deux vecteurs de n observations X et Y [4].

2.3.3 Product-moment correlation

Dans la littérature anglo-saxonne, on parle souvent de "product-moment correlation" à propos du coefficient de corrélation de Pearson. Cela s'explique par le fait qu'il peut s'exprimer comme la moyenne du produit des variables centrées réduites. Si l'on désigne par $\overset{cr}{x}$ (resp. $\overset{cr}{y}$) les valeurs de X (resp. Y) centrées et réduites c.-à-d.

$$\overset{cr}{x_i} = \frac{x_i - \bar{x}}{s_x}$$

Le coefficient de corrélation empirique peut s'écrire

$$\hat{r} = \frac{1}{n} \sum_{i=1}^{n} \overset{cr}{x_i} \times \overset{cr}{y_i} \tag{2.10}$$

En particulier, lorsque les données sont centrées et réduites, covariance et corrélation empiriques sont équivalents.

2.3.4 Biais et Coefficient de corrélation ajusté

Le coefficient de corrélation empirique est un estimateur biaisé. Fort heureusement, le biais devient négligeable lorsque l'effectif augmente. L'espérance de l'estimateur s'écrit ([1], page 107) :

$$E[\hat{r}] = r - \frac{r(1 - r^2)}{2n}$$

3. Voir `http://fr.wikipedia.org/wiki/Régression_linéaire_multiple`
4. Voir `http://en.wikipedia.org/wiki/Correlation_coefficient`

Pour cette raison, certains logiciels proposent un coefficient de corrélation ajusté [5] ([6], page 274)

$$\hat{r}_{aj} = \sqrt{1 - \frac{n-1}{n-2}(1-\hat{r}^2)} \tag{2.11}$$

Bien entendu, l'ajustement est d'autant plus sensible que l'effectif est faible. Lorsque n est élevé, $\hat{r}$ et $\hat{r}_{aj}$ se confondent.

2.3.5 Exemples numériques

Détails des calculs sur un exemple. Reprenons les variables cylindrée (X) et puissance (Y) de notre fichier "voitures". Nous détaillons les calculs dans la feuille EXCEL (Figure 2.4) :

	A	B	C	D	E	F	G
1	Numero	Modele	Cylindree	Puissance	XY	X²	Y²
2	1	Daihatsu Cuore	846	32	27072	715716	1024
3	2	Suzuki Swift 1.0 GLS	993	39	38727	986049	1521
4	3	Fiat Panda Mambo L	899	29	26071	808201	841
5	4	VW Polo 1.4 60	1390	44	61160	1932100	1936
6	5	Opel Corsa 1.2i Eco	1195	33	39435	1428025	1089
7	6	Subaru Vivio 4WD	658	32	21056	432964	1024
8	7	Toyota Corolla	1331	55	73205	1771561	3025
9	8	Opel Astra 1.6i 16V	1597	74	118178	2550409	5476
10	9	Peugeot 306 XS 108	1761	74	130314	3101121	5476
11	10	Renault Safrane 2.2. V	2165	101	218665	4687225	10201
12	11	Seat Ibiza 2.0 GTI	1983	85	168555	3932289	7225
13	12	VW Golt 2.0 GTI	1984	85	168640	3936256	7225
14	13	Citroen ZX Volcane	1998	89	177822	3992004	7921
15	14	Fiat Tempra 1.6 Liberty	1580	65	102700	2496400	4225
16	15	Fort Escort 1.4i PT	1390	54	75060	1932100	2916
17	16	Honda Civic Joker 1.4	1396	66	92136	1948816	4356
18	17	Volvo 850 2.5	2435	106	258110	5929225	11236
19	18	Ford Fiesta 1.2 Zetec	1242	55	68310	1542564	3025
20	19	Hyundai Sonata 3000	2972	107	318004	8832784	11449
21	20	Lancia K 3.0 LS	2958	150	443700	8749764	22500
22	21	Mazda Hachtback V	2497	122	304634	6235009	14884
23	22	Mitsubishi Galant	1998	66	131868	3992004	4356
24	23	Opel Omega 2.5i V6	2496	125	312000	6230016	15625
25	24	Peugeot 806 2.0	1998	89	177822	3992004	7921
26	25	Nissan Primera 2.0	1997	92	183724	3988009	8464
27	26	Seat Alhambra 2.0	1984	85	168640	3936256	7225
28	27	Toyota Previa salon	2438	97	236486	5943844	9409
29	28	Volvo 960 Kombi aut	2473	125	309125	6115729	15625
30	**n**		**Moyenne**			**Somme**	
31	28		1809.07	77.71	4451219	102138444	197200
32							
33			Numérateur	514679.571			
34			Dénominateur	543169.291			
35			**Corrélation**	**0.9475**			
36							
37			Coef.Corr.Excel	0.9475			

Fig. 2.4. Détails des calculs - Estimation de la corrélation

- Au bout des colonnes C et D, nous disposons toujours des moyennes empiriques.
- Nous formons les quantités $(x_i y_i)$, x_i^2 et y_i^2. Nous calculons leurs sommes respectives : 4451219, 102138444 et 197200.

5. Voir le parallèle avec le coefficient de détermination ajusté en régression linéaire multiple `http://fr.wikipedia.org/wiki/Régression_linéaire_multiple`

- A partir de la formule 2.9, nous obtenons le numérateur = 514679.571 et le dénominateur = 543169.291.
- Reste à former le rapport, la corrélation entre la cylindrée et la puissance est $\hat{r} = 0.9475$.
- La fonction "COEFFICIENT.CORRELATION(...)" du tableur EXCEL propose la même valeur.

Nuage de points. Il y a une forte liaison linéaire entre "cylindrée" et "puissance", ce que confirme le graphique nuage de points (Figure 2.5). On notera aussi, et le coefficient de corrélation ne sait pas traduire ces informations, que 2 points semblent s'écarter des autres, mais pas de la même manière :

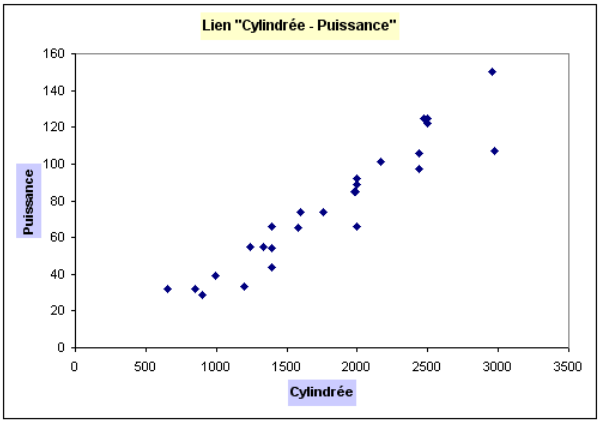


Fig. 2.5. Nuage de points "Cylindrée vs. Puissance"

- La "Lancia K 3.0 LS" est une grosse cylindrée, très puissante. Elle s'écarte du nuage certes, mais elle est dans la lignée de la liaison entre les deux variables.
- La "Hyundai Sonata 3000" est aussi une grosse cylindrée, mais elle est relativement anémique. Le point est un peu à l'écart des autres, tout comme la Lancia, mais elle ne respecte pas, apparemment, l'apparente liaison (visuelle et numérique) entre cylindrée et puissance. Si on retire cette observation, la corrélation est renforcée, elle passe à 0.9635.

Comparaison de coefficients de corrélation. Maintenant, nous pouvons comparer les coefficients de corrélation calculés sur différentes variables. Reprenons notre exemple des voitures, calculons le coefficient de corrélation de consommation avec les autres variables, nous obtenons respectivement : cylindrée = 0.892, puissance = 0.888 et poids = 0.926.

La variable "consommation" est singulièrement corrélée avec l'ensemble des variables. Le lien avec poids semble plus élevé que le lien avec puissance. *Mais sans l'arsenal de l'inférence statistique, nous ne pouvons pas affirmer s'il est significativement plus élevé que les autres.*

2.4 Test de significativité

2.4.1 Spécifications du test

Le premier test qui vient à l'esprit est la significativité de la corrélation c.-à-d. le coefficient de corrélation est-il significativement différent de 0 ?

Le test s'écrit :

$$H_0 : r = 0$$
$$H_1 : r \neq 0$$

Remarque 2 (Autres hypothèses alternatives). On peut vouloir définir une hypothèse alternative différente ($H_1 : r < 0$ ou $H_1 : r > 0$). Les caractéristiques des distributions restent les mêmes. Pour un risque α donné, seul est modifié le seuil de rejet de H_0 puisque le test est unilatéral dans ce cas.

Test exact. Le test étudié dans cette section est paramétrique. On suppose *a priori* que le couple (X, Y) suit une loi normale bivariée[6]. Dans ce cas : la distribution sous H_0 de la statistique du test que nous présenterons plus bas est exacte ; le test de significativité équivaut à un test d'indépendance.

Test asymptotique. Cette restriction est moins contraignante lorsque n est suffisamment grand[7]. A partir de 25 observations, l'approximation est bonne, même si nous nous écartons (un peu) de la distribution normale conjointe ([10], page 308). La distribution est asymptotiquement valable sous l'hypothèse $r = 0$. Mais le test de significativité revient simplement à tester l'absence ou la présence de corrélation.

Statistique du test. Sous H_0, la statistique :

$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-2}}} \tag{2.12}$$

suit une loi de Student à $(n-2)$ degrés de liberté.

Région critique. La région critique (rejet de l'hypothèse nulle) du test au risque α s'écrit :

$$R.C. : |t| > t_{1-\frac{\alpha}{2}}(n-2)$$

où $t_{1-\frac{\alpha}{2}}(n-2)$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi de Student à $(n-2)$ degrés de liberté. Il s'agit d'un test bilatéral.

Probabilité critique (p-value). Plutôt que de comparer la statistique calculée avec la seuil théorique fournie par la loi de Student, les logiciels proposent souvent la probabilité critique (*p-value*) que l'on doit comparer au risque α que l'on s'est fixé. Si la p-value est plus petite, alors nous rejetons l'hypothèse nulle.

6. Si (X, Y) suit une loi normale bivariée, alors X et Y suivent individuellement une loi normale. En revanche, ce n'est pas parce que X et Y sont individuellement gaussiens que le couple (X, Y) l'est forcément. Enfin, si X **ou** Y n'est pas gaussien, le couple (X, Y) ne l'est pas non plus.

7. Voir `http://faculty.vassar.edu/lowry/ch4pt1.html` et `http://www2.chass.ncsu.edu/garson/PA765/correl.htm#assume`

2.4.2 Un exemple numérique

Reprenons le calcul de la corrélation entre la cylindrée et la puissance (Figure 2.4). Nous souhaitons tester sa significativité au risque $\alpha = 0.05$. Nous avions $n = 28$, et $\hat{r} = 0.9475$.

Nous devons calculer les éléments suivants :

- La statistique du test $t = \frac{0.9475}{\sqrt{\frac{1-0.9475^2}{28-2}}} = 15.1171$
- Le seuil théorique au risque $\alpha = 0.05$ est $t_{0.975}(28-2) = 2.0555$
- Nous concluons donc au rejet de l'hypothèse nulle c.-à-d. les résultats que nous obtenons à partir des données ne sont pas compatibles avec une absence de corrélation. On s'en serait douté avec une valeur aussi élevée. A la différence que maintenant, nous pouvons associer un risque à la prise de décision.

2.4.3 Test asymptotique (bis)

De manière générale, $\hat{r}$ tend lentement vers la loi normale. Quand $n \to +\infty$, t suit une loi de Student à degrés de liberté infini, donc vers la loi normale.

Sous l'hypothèse $H_0 : r = 0$, la convergence est plus rapide. Lorsque $n > 100$, la loi de $\hat{r}$ peut être approximée à l'aide de la loi normale $\mathcal{N}(0; \frac{1}{\sqrt{n-1}})$. Le test de significativité peut s'appuyer sur cette distribution.

2.5 Autres tests et intervalle de confiance

Pour calculer un intervalle de confiance ou tester la conformité de r avec une autre valeur que 0, il faudrait connaître la distribution de la statistique de manière générique c.-à-d. quelle que soit la vraie valeur de r dans la population.

Or, on se rend compte que dans un voisinage autre que $r = 0$, la convergence vers la loi normale est plus lente et, pour les petits effectifs, la distribution de $\hat{r}$ tend à être dissymétrique à gauche ([2], page 15).

Pour remédier à cela, il est conseillé de passer par une transformation dite de Fisher.

2.5.1 Transformation de Fisher

La transformation de Fisher s'écrit

$$\hat{z} = \frac{1}{2} \ln \frac{1+\hat{r}}{1-\hat{r}} \tag{2.13}$$

Elle est distribuée asymptotiquement selon une **loi normale** de paramètres [8]

8. Il existe une approximation ([1], page 108) plus précise de l'espérance $E[\hat{z}] \approx \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{r}{2(n-1)}$. Il y a un léger biais, mais il devient très vite négligeable dès que n augmente.

$$E[\hat{z}] \approx \frac{1}{2}\ln\frac{1+r}{1-r}$$
$$V[\hat{z}] \approx \frac{1}{n-3}$$

L'approximation est bonne dès les (relativement) petites valeurs de n (dès $n > 10$ en pratique).

Nous pouvons nous appuyer sur cette statistique pour réaliser le test de significativité ci-dessus. Mais, plus intéressant encore, la transformation nous offre d'autres possibilités.

2.5.2 Intervalle de confiance

Nous pouvons calculer un intervalle de confiance pour $\hat{r}$. Il faut pour cela garder à l'esprit que l'on peut obtenir $\hat{r}$ à partir de $\hat{z}$ en utilisant la relation

$$\hat{r} = \frac{e^{2\hat{z}} - 1}{e^{2\hat{z}} + 1} \tag{2.14}$$

Voici la démarche à adopter pour obtenir l'intervalle de confiance au niveau de confiance $(1-\alpha)$:

- Calculer $\hat{z}$ à partir de $\hat{r}$ (Equation 2.13)
- Calculer les bornes de l'intervalle de confiance de z avec

$$z_{1,2} = \hat{z} \pm u_{1-\frac{\alpha}{2}} \times \sqrt{\frac{1}{n-3}} \tag{2.15}$$

- En déduire alors les bornes de l'intervalle de confiance de r (Equation 2.14)

Exemple numérique. Nous souhaitons calculer l'intervalle de confiance de la corrélation entre cylindrée et puissance pour un niveau de confiance de 95%. Rappelons que $\hat{r} = 0.9475$.

- Le quantile de la loi normale centrée réduite d'ordre 0.975 est $u_{0.975} = 1.96$
- La transformation de Fisher nous donne $\hat{z} = \frac{1}{2}\ln\frac{1+0.9475}{1-0.9475} = 1.8072$
- L'écart type de $\hat{z}$ est égal à $\sqrt{\frac{1}{28-3}} = 0.2$
- La borne basse de l'intervalle de confiance s'écrit $z_1 = 1.8072 - 1.96 \times 0.2 = 1.4152$; selon le même procédé, la borne haute $z_2 = 2.1992$
- Nous en déduisons les bornes de l'intervalle de confiance du coefficient de corrélation :

$$r_1 = \frac{e^{2\times 1.4152} - 1}{e^{2\times 1.4152} + 1} = 0.8886$$
$$r_2 = \frac{e^{2\times 2.1992} - 1}{e^{2\times 2.1992} + 1} = 0.9757$$

L'intervalle de confiance au niveau 95% de la corrélation entre la cylindrée et la puissance est

$$[0.8886\,;\,0.9757]$$

2.5.3 Comparaison à un standard (autre que 0)

La transformation nous permet d'aller plus loin que le simple test de significativité, nous avons la possibilité de comparer la valeur du coefficient de corrélation avec une valeur de référence r_0. La loi associée à z est valable quelle que soit la valeur de r dans la population parente.

Nous passons par la transformation de Fisher, avec $z_0 = \frac{1}{2}\ln\frac{1+r_0}{1-r_0}$, l'hypothèse nulle du test s'écrit

$$H_0 : z = z_0$$

La statistique du test U est

$$U = \frac{\hat{z} - z_0}{\sqrt{\frac{1}{n-3}}} = (\hat{z} - z_0) \times \sqrt{n-3} \tag{2.16}$$

Elle suit une loi normale centrée réduite.

Exemple : Corrélation cylindrée - puissance. Nous souhaitons effectuer le test **unilatéral** suivant au risque 5%

$$H_0 : r = 0.9$$
$$H_1 : r > 0.9$$

Les étapes du calcul sont les suivantes

- Nous calculons la valeur de référence transformée $z_0 = \frac{1}{2}\ln\frac{1+0.9}{1-0.9} = 1.4722$
- Rappelons que $\hat{r} = 0.9475$ et $\hat{z} = 1.8072$
- La statistique du test est $U = (\hat{z} - z_0) \times \sqrt{n-3} = (1.8072 - 1.4722) \times \sqrt{28-3} = 1.6750$
- Que nous devons comparer avec le quantile d'ordre $1-\alpha = 1-0.05 = 0.95$ de la loi normale centrée réduite c.-à-d. $u_{0.95} = 1.6449$
- Au risque $\alpha = 5\%$, l'hypothèse nulle n'est pas compatible avec nos données, nous acceptons H_1

2.5.4 Comparaison de 2 coefficients de corrélation (échantillons indépendants)

Autre possibilité qu'introduit la transformation de Fisher : la comparaison les corrélations dans deux populations différentes. Mettons que nous souhaitons comparer la corrélation entre le poids et la taille chez les hommes et chez les femmes. Est-ce qu'elle est identique dans les deux populations ?

Nous travaillons sur 2 échantillons indépendants, extraits au hasard dans chaque sous population. La corrélation théorique est r_1 (resp. r_2) chez les femmes (resp. chez les hommes). Le test d'hypothèses s'écrit :

$$H_0 : r_1 = r_2$$
$$H_1 : r_1 \neq r_2$$

Nous disposons de 2 échantillons de taille n_1 et n_2. Nous introduisons la statistique

$$D = \hat{z}_1 - \hat{z}_2 \tag{2.17}$$

Sous H_0, puisque les estimateurs $\hat{r}$ (et par conséquent $\hat{z}$) sont indépendants (estimés sur des échantillons indépendants), la statique D suit asymptotiquement une loi normale de paramètres

$$\begin{aligned} E[D] &= 0 \\ V[D] &= \frac{1}{n_1 - 3} + \frac{1}{n_2 - 3} \end{aligned}$$

Au risque α, la région critique du test bilatéral s'écrit :

$$R.C. : U = \frac{|\hat{z}_1 - \hat{z}_2|}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}} \geq u_{1-\frac{\alpha}{2}}$$

Exemple numérique : comparer la corrélation taille - poids chez les hommes et chez les femmes. Nous disposons d'un échantillon de $n_1 = 15$ femmes, et $n_2 = 20$ hommes (Figure 2.6). Nous souhaitons tester l'égalité du coefficient de corrélation entre le poids et la taille dans les deux sous-populations au risque de 5%. Les étapes du calcul sont énumérées ci-dessous.

- Nous calculons les coefficients de corrélation, nous obtenons $\hat{r}_1 = 0.5661$ et $\hat{r}_2 = 0.4909$
- Nous appliquons la transformation de Fisher, $\hat{z}_1 = 0.6417$ et $\hat{z}_2 = 0.5372$
- Nous calculons la statistique $D = \hat{z}_1 - \hat{z}_2 = 0.1045$, puis sa variance $V(D) = \frac{1}{15-3} + \frac{1}{20-3} = 0.1422$
- Nous en déduisons alors $U = \frac{|0.1045|}{\sqrt{0.1422}} = \frac{0.3652}{0.3770} = 0.2771$
- Que nous comparons au quantile d'ordre 0.975 de la loi normale centrée réduite, soit $u_{0.975} = 1.96$
- Conclusion : au risque de 5%, les données sont compatibles avec l'hypothèse nulle c.-à-d. le coefficient de corrélation entre le poids et taille n'est pas significativement différent chez les hommes et les femmes.

2.5.5 Comparaison de K ($K \geq 2$) coefficients (échantillons indépendants)

Il est possible de généraliser ce test pour comparer K coefficients de corrélation dans K sous-populations. La statistique du test s'écrit différemment, elle suit une loi du χ^2 dans ce cas (voir [2], page 22). Il s'agit bien souvent de comparer le même coefficient de corrélation sur plusieurs sous-populations.

Remarque 3 (C'est une vraie généralisation). Lorsque $K = 2$, nous devrions retrouver le test précédent, nous vérifierons cela sur le même exemple que précédemment (section 2.5.4).

L'hypothèse nulle du test est

$$H_0 : r_1 = r_2 = \cdots = r_K$$

L'hypothèse alternative est "un des coefficients au moins s'écarte des autres".

		Femmes			Hommes	
		Poids (kg)	Taille (m)		Poids (kg)	Taille (m)
1	1	77.56	1.70	1	83.91	1.65
2	2	50.35	1.54	2	68.95	1.74
3	3	76.66	1.63	3	100.70	1.76
4	4	62.60	1.63	4	81.19	1.71
5	5	58.07	1.44	5	93.44	1.81
6	6	72.57	1.68	6	87.54	1.71
7	7	92.53	1.64	7	97.52	1.78
8	8	76.66	1.63	8	78.02	1.69
9	9	58.06	1.54	9	81.65	1.77
10	10	71.67	1.54	10	81.65	1.65
11	11	68.04	1.62	11	79.83	1.61
12	12	70.06	1.58	12	90.72	1.77
13	13	61.69	1.56	13	75.75	1.73
14	14	67.59	1.50	14	83.91	1.69
15	15	59.87	1.62	15	71.23	1.75
16				16	99.79	1.87
17				17	73.94	1.74
18				18	76.20	1.69
19				19	115.21	1.78
20				20	70.31	1.73

n1	15	
r1	0.5661	
z1	0.6417	

n2	20	
r2	0.4909	
z2	0.5372	

D	0.1045
V(D)	0.1422

U	0.2771
u(0.975)	1.96

Fig. 2.6. Comparaison de 2 coefficients de corrélation - Échantillons indépendants

La statistique du test s'écrit :

$$\chi^2 = \sum_{k=1}^{K}(n_k - 3)\hat{z}_k^2 - \frac{[\sum_{k=1}^{K}(n_k - 3)\hat{z}_k]^2}{\sum_{k=1}^{K}(n_k - 3)} \tag{2.18}$$

où n_k est l'effectif de l'échantillon ayant servi à mesurer la corrélation $\hat{r}_k$; $\hat{z}_k$ est la transformation de Fisher de $\hat{r}_k$ c.-à-d. $\hat{z}_k = \frac{1}{2}\ln\frac{1+\hat{r}_k}{1-\hat{r}_k}$.

Sous H_0, la statistique du test suit une loi du $\chi^2(K-1)$ à $K-1$ degrés de liberté. On rejette l'hypothèse nulle lorsqu'elle est supérieure au quantile $\chi^2_{1-\alpha}(K-1)$ de la loi théorique pour un risque α.

Exemple numérique 1 : comparaison de la corrélation poids vs. consommation des véhicules de différentes origines. Nous souhaitons vérifier, au risque de 5%, que la corrélation entre le poids et la consommation des véhicules est la même pour des véhicules en provenance de l'Europe (France, Allemagne, etc.), du Japon, et des USA. Le fichier est disponible sur le site DASL (Data and Story Library) [9]. Du fichier original, nous avons supprimé l'observation atypique (la fameuse Buick Estate Wagon). Nous disposons pour chaque catégorie de véhicule de $n_1 = 9$, $n_2 = 7$ et $n_3 = 21$ observations.

Tous les calculs ont été menés dans une feuille EXCEL (Figure 2.7), en voici les détails :

9. http://lib.stat.cmu.edu/DASL/Stories/FuelEfficientBuickWagon.html

Europe		Japon		U.S.A	
Poids	Conso	Poids	Conso	Poids	Conso
3.41	14.52	2.56	8.55	4.05	15.18
2.83	11.59	2.30	8.65	3.61	12.25
1.99	7.47	1.98	6.90	3.94	12.72
2.19	7.71	1.92	6.70	2.16	7.84
2.60	10.94	2.14	7.97	2.23	7.61
1.93	7.37	2.02	7.40	3.38	11.42
2.13	6.31	2.82	10.69	3.07	11.31
3.14	13.84			3.62	12.65
2.80	10.89			3.41	13.00
				3.84	13.84
				3.73	13.37
				3.96	14.26
				3.83	12.93
				2.59	8.88
				2.91	10.74
				2.67	8.59
				2.67	8.28
				2.60	8.17
				2.70	8.78
				2.56	7.02
				2.20	6.88

r1	0.9716	**r2**	0.9540	**r3**	0.9647
n1	9	**n2**	7	**n3**	21

z1	2.1198	**z2**	1.8741	**z3**	2.0092
n1-3	6	**n2-3**	4	**n3-3**	18

A	3178.7259
B	28
C	113.6718

KHI-2	0.1459
ddl	2

KHI-2 (0.95; 2)	5.9915

p-value	0.9297

Fig. 2.7. Comparaison de $K = 3$ coefficients de corrélation - Échantillons indépendants

- Pour chaque origine des véhicules, nous disposons des deux colonnes de données (Poids et Consommation).
- Nous obtenons les coefficients de corrélation empiriques $\hat{r}_1 = 0.9716$, $\hat{r}_2 = 0.9540$, $\hat{r}_3 = 0.9647$; en appliquant la transformation de Fisher, nous avons : $\hat{z}_1 = 2.1198$, $\hat{z}_2 = 1.8741$, $\hat{z}_3 = 2.0092$.
- Nous formons alors $A = \sum_k (n_k - 3)\hat{z}_k = 3178.7259$; $B = \sum_k (n_k - 3) = 28$; $C = \sum_k (n_k - 3)\hat{z}_k^2 = 113.6718$.
- La statistique du test est $\chi^2 = C - \frac{A}{B} = 0.1459$.
- Le quantile d'ordre $1-\alpha = 95\%$ de la loi du χ^2 à $(K-1) = 2$ degrés de liberté est $\chi^2_{0.95}(2) = 5.9915$. Nos données sont compatibles avec l'hypothèse nulle : les corrélations sont les mêmes quelle que soit l'origine des véhicules.
- De la même manière, nous aurions pu calculer la probabilité critique du test (la p-value), elle est égale à 0.9297, largement supérieure au risque 5%. La conclusion est bien évidemment la même.

Exemple numérique 2 : Comparaison de la corrélation taille - poids chez les hommes et chez les femmes. Le test est une généralisation de la comparaison de 2 coefficients. Vérifions que

les résultats sont en accord avec notre exemple de la section 2.5.4. Détaillons de nouveaux les calculs en reprenant les notations de l'exemple précédent

- $A = [(15-3) \times 0.6417 + (20-3) \times 0.5372]^2 = 283.3678$
- $B = (15-3) + (20-3) = 29$
- $C = (15-3) \times 0.6417^2 + (20-3) \times 0.5372^2 = 9.8481$
- Ainsi, la statistique du test est $\chi^2 = C - \frac{A}{B} = 0.0768$, que l'on comparera à $\chi^2_{0.95}(1) = 3.8415$. Conformément au test précédent, on conclut, au risque 5%, que les données sont compatibles avec l'hypothèse d'égalité des coefficients de corrélation.
- En regardant de plus près les résultats, nous constatons que $\sqrt{0.0768} = 0.2771$. On retrouve exactement la valeur de la statistique du test basé sur la loi normale. Ce n'est guère étonnant, en effet n'oublions pas qu'il y a une relation entre la loi normale et la loi du χ^2 à 1 degré de liberté c.-à-d. $[\mathcal{N}(0;1)]^2 \equiv \chi^2(1)$. Les deux tests sont totalement équivalents.

2.5.6 Comparaison de 2 coefficients de corrélation (même échantillon) - Cas 1

Autre analyse intéressante dans la pratique, nous souhaitons comparer les corrélations respectives de deux variables X et Z avec la variable Y. La situation est un peu plus complexe car les corrélations sont calculées sur un seul et même échantillon.

L'hypothèse nulle du test est naturellement

$$H_0 : r_{yx} = r_{yz}$$

On peut vouloir construire un test unilatéral ($r_{yx} > r_{yz}$ ou $r_{yx} < r_{yz}$) ou bilatéral ($r_{yx} \neq r_{yz}$).

Dans ce cadre, le test t de Williams est conseillé dès lors que n est assez grand ($n \geq 20$). La statistique s'écrit ([2], page 24)

$$t = (\hat{r}_{yx} - \hat{r}_{yz}) \sqrt{\frac{(n-1)(1+\hat{r}_{xz})}{2\frac{n-1}{n-3}|R| + \bar{r}^2(1-\hat{r}_{xz})^3}} \tag{2.19}$$

où $\bar{r} = (\hat{r}_{yx} + \hat{r}_{yz})/2$; $|R| = 1 - \hat{r}^2_{yx} - \hat{r}^2_{yz} - \hat{r}^2_{xz} + 2\hat{r}_{yx}\hat{r}_{yz}\hat{r}_{xz}$

t suit une loi de Student à $(n-3)$ degrés de liberté.

Remarque 4 (X et Z sont orthogonaux). Nous remarquons que le degré du lien entre les variables X et Z influe sur les résultats. Si X et Z sont orthogonaux (c.-à-d. $r_{xz} = 0$), la statistique dépend uniquement des corrélations r_{yx} et r_{yz}.

Exemple numérique : comparaison de la corrélation "consommation - puissance et consommation - cylindrée. Reprenons notre fichier des voitures (Figure 2.2). Nous souhaitons savoir si, à 5%, la corrélation de la consommation (Y) avec la cylindrée (la taille du moteur, X) est comparable à sa corrélation avec la puissance (Z). Nous sommes sur un test bilatéral, on veut vérifier si l'écart observé est statistiquement significatif.

Conformément à la formule 2.19, nous construisons la feuille EXCEL (Figure 2.8) :

	A	B	C	D	E	F	G	H
1			X	Z	Y			
2	Numero	Modele	Cylindree	Puissance	Conso			
3	1	Daihatsu Cuore	846	32	5.7			
4	2	Suzuki Swift 1.0 GLS	993	39	5.8		n	28
5	3	Fiat Panda Mambo L	899	29	6.1			
6	4	VW Polo 1.4 60	1390	44	6.5		r (Y,X)	0.8919
7	5	Opel Corsa 1.2i Eco	1195	33	6.8		r (Y,Z)	0.8878
8	6	Subaru Vivio 4WD	658	32	6.8			
9	7	Toyota Corolla	1331	55	7.1		r (X,Z)	0.9475
10	8	Opel Astra 1.6i 16V	1597	74	7.4			
11	9	Peugeot 306 XS 108	1761	74	9.0		A	0.0041
12	10	Renault Safrane 2.2. V	2165	101	11.7		B	52.5838
13	11	Seat Ibiza 2.0 GTI	1983	85	9.5		R	0.0191
14	12	VW Golt 2.0 GTI	1984	85	9.5		r-barre	0.8898
15	13	Citroen ZX Volcane	1998	89	8.8		C	0.0001
16	14	Fiat Tempra 1.6 Liberty	1580	65	9.3			
17	15	Fort Escort 1.4i PT	1390	54	8.6		t	0.1448
18	16	Honda Civic Joker 1.4	1396	66	7.7		ddl	25
19	17	Volvo 850 2.5	2435	106	10.8			
20	18	Ford Fiesta 1.2 Zetec	1242	55	6.6		t(0.975 ; 25)	2.0595
21	19	Hyundai Sonata 3000	2972	107	11.7			
22	20	Lancia K 3.0 LS	2958	150	11.9		p-value	0.8861
23	21	Mazda Hachtback V	2497	122	10.8			
24	22	Mitsubishi Galant	1998	66	7.6			
25	23	Opel Omega 2.5i V6	2496	125	11.3			
26	24	Peugeot 806 2.0	1998	89	10.8			
27	25	Nissan Primera 2.0	1997	92	9.2			
28	26	Seat Alhambra 2.0	1984	85	11.6			
29	27	Toyota Previa salon	2438	97	12.8			
30	28	Volvo 960 Kombi aut	2473	125	12.7			

Fig. 2.8. Comparaison de 2 corrélations du même échantillon - Cas 1

- Notre effectif est $n = 28$.
- Nous calculons les corrélations à comparer $\hat{r}_{yx} = 0.8919$ et $\hat{r}_{yz} = 0.8878$. Nous voulons savoir si l'écart observé est significatif c.-à-d. transposable dans la population (H_1) ou uniquement du aux fluctuations d'échantillonnage (H_0).
- Nous calculons la corrélation $\hat{r}_{xz} = 0.9475$. Nous constatons qu'elles sont très liées. Peut être d'ailleurs qu'elles amènent le même type d'information vis à vis de Y, nous vérifierons cette assertion dans la partie de ce support consacrée aux corrélation partielles.
- Nous calculons l'écart $A = (\hat{r}_{yx} - \hat{r}_{yz}) = 0.0041$
- $B = (n-1)(1 + \hat{r}_{xz} = 52.5838$
- $|R| = 1 - \hat{r}_{yx}^2 - \hat{r}_{yz}^2 - \hat{r}_{xz}^2 + 2\hat{r}_{yx}\hat{r}_{yz}\hat{r}_{xz} = 0.0191$
- $\bar{r} = (\hat{r}_{yx} + \hat{r}_{yz})/2 = 0.8898$
- $C = (1 - \hat{r}_{xz})^3 = 0.0001$
- Nous obtenons la statistique du test $t = A\sqrt{\frac{B}{2\frac{27}{25}0.0191 + 0.8898 \times 0.0001}} = 0.1448$
- Que nous comparons au seuil critique $\mathcal{T}_{0.975}(25) = 2.0595$.
- Au risque 5%, nos données sont compatibles avec l'hypothèse nulle, la consommation est identiquement corrélée à la cylindrée et à la puissance.
- La p-value du test égal à 0.8861 conduit bien évidemment à la même conclusion.

2.5.7 Comparaison de 2 coefficients de corrélation (même échantillon) - Cas 2

Toujours à partir sur un même échantillon, ce second test consiste à opposer

$$H_0 : r_{xy} = r_{zw}$$
$$H_1 : r_{xy} \neq r_{zw}$$

Le test peut être unilatéral (c.-à-d. $H_1 : r_{xy} < r_{zw}$ ou $r_{xy} > r_{zw}$).

De prime abord, ce test paraît assez étrange. Est-ce que comparer des corrélations calculées sur des concepts différents a réellement un sens ? Prenons l'exemple des voitures, opposer la corrélation entre la puissance et la consommation, d'une part, et la corrélation entre le poids et le prix, d'autre part, ne paraît pas très pertinent.

On comprend mieux le sens de ce test à la lumière de l'exemple proposé par une des rares références qui le décrit (voir [2], page 24). Pour un ensemble d'électeurs, on calcule la corrélation entre les donations et les intentions de votes, une année donnée, puis 4 ans plus tard. L'objectif est de vérifier si le lien entre ces deux variables a été modifié entre temps.

De cet exemple, nous retiendrons avant tout l'idée d'**appariement**. Nous voulons comparer l'intensité d'un lien avant et après l'occurrence d'un évènement, qui peut être simplement un certain délai, mais qui peut être aussi une action particulière. Mais la notion d'appariement est plus large. Il y a effectivement la situation "avant - après". Mais nous pouvons la définir surtout comme des mesures effectuées sur une unité statistique : dans un ménage, mesurer et comparer une caractéristique chez l'homme et la femme ; comparer la même variable chez des jumeaux ; etc. [10].

Le test de Clark et Dunn est conseillée pour cette configuration. Il suit asymptotiquement une loi normale centrée réduite, il est valable dès lors que $n \geq 20$. Par commodités, nous numéroterons les variables $X = 1$, $Y = 2$, $Z = 3$ et $W = 4$. Nous écrirons par exemple $\hat{r}_{12}$ pour $\hat{r}_{xy}$, ou $\hat{r}_{34}$ pour $\hat{r}_{zw}$, etc.

La statistique du test s'écrit

$$U = (\hat{z}_{12} - \hat{z}_{34})\sqrt{\frac{n-3}{2-2\bar{s}}} \tag{2.20}$$

avec

- $\hat{z} = \frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}}$, la transformation de Fisher ;
- $\bar{s} = \frac{\psi}{(1-\bar{r}^2)^2}$;
- $\bar{r} = \frac{\hat{r}_{12}+\hat{r}_{34}}{2}$;
- $\psi = 0.5\{[(\hat{r}_{13} - \hat{r}_{23}\bar{r})(\hat{r}_{24} - \hat{r}_{23}\bar{r})] + [(\hat{r}_{14} - \hat{r}_{13}\bar{r})(\hat{r}_{23} - \hat{r}_{13}\bar{r})] + [(\hat{r}_{13} - \hat{r}_{14}\bar{r})(\hat{r}_{24} - \hat{r}_{14}\bar{r})] + [(\hat{r}_{14} - \hat{r}_{24}\bar{r})(\hat{r}_{23} - \hat{r}_{24}\bar{r})]\}$

Exemple : les donations au parti. Reprenons directement l'exemple décrit dans l'ouvrage de Chen et Popovich ([2], page 25). Il s'agit de tester, pour $n = 203$ votants, si le lien entre les donations au parti et les intentions de vote a évolué dans un laps de temps de 4 années. Les corrélations à comparer sont $\hat{r}_{12} = 0.3$ et $\hat{r}_{34} = 0.4$.

Nous disposons des corrélations croisées : $\hat{r}_{13} = 0.6$, $\hat{r}_{14} = 0.2$, $\hat{r}_{23} = 0.3$, $\hat{r}_{24} = 0.7$.

A partir des équations ci-dessus, nous obtenons $\bar{r} = 0.35$, $\psi = 0.3125$ et $\bar{s} = 0.4059$.

10. Voir `http://www.tufts.edu/~gdallal/paired.htm`

La statistique du test est égal à $U = -1.48$. Au risque 5%, pour un test bilatéral, nous comparons $|U| = 1.48$ avec le quantile de la loi normale centrée réduite $u_{0.975} = 1.96$. Les données sont compatibles avec l'hypothèse nulle, 4 années plus tard, le lien entre les intentions de vote et les donations n'a pas évolué significativement.

Commentaires sur la comparaison des coefficients de corrélations. Comme le notent Chen et Popovich dans leur ouvrage ([2], page 25), les tests de comparaison de coefficients de corrélations sont peu décrits, peu répandus, et de ce fait rarement disponibles dans les logiciels (à moins que ce ne soit l'inverse, c'est parce qu'ils sont peu programmés qu'ils sont peu utilisés). C'est regrettable car les applications pratiques sont nombreuses, elles ouvrent d'autres pistes pour l'exploration des données. De plus, argument important qui milite en faveur de leur diffusion, le dispositif est très souple : les tests restent valables pour les mesures de corrélation dérivées du coefficient de Pearson, mesures que nous décrirons dans le chapitre 3 de ce support.

2.6 Problèmes et cas pathologiques

"Corrélation n'est pas causalité". C'est une phrase maintes fois répétée dans tous les ouvrages. En effet, le coefficient de corrélation est un indicateur statistique, avec ses forces et ses faiblesses. Il ne faut surtout pas en faire une référence absolue. Il importe de délimiter clairement son champ d'action et identifier les cas où ses indications sont sujettes à caution. La qualité des interprétations consécutives aux calculs en dépend (voir aussi [3], pages 93-94, concernant les "petites corrélations").

2.6.1 Corrélation fortuite

La corrélation peut parfois être totalement fortuite. Johnston ([4], page 10) rapporte par exemple que sur les données annuelles de 1897 à 1985, des études ont montré une corrélation de 0.91 entre le revenu national américain et le nombre de tâches solaires (les zones sombres du soleil, ce sont des zones moins chaudes). Personne ne peut décemment soutenir qu'il y a une relation quelconque entre ces 2 grandeurs.

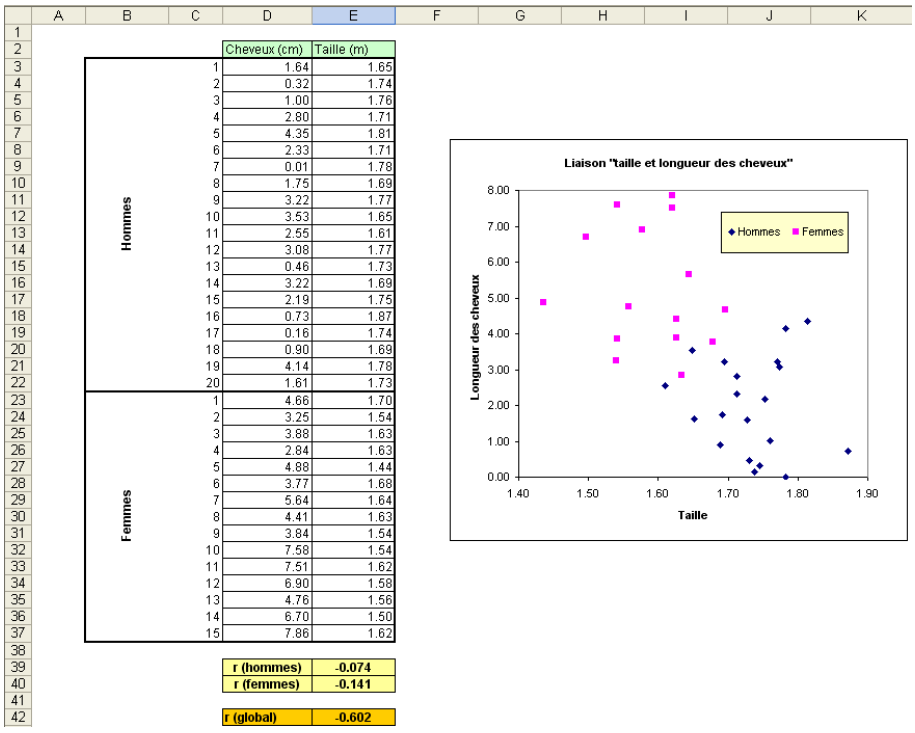
2.6.2 Facteur confondant

La corrélation peut aussi cacher l'influence d'un autre facteur. On montre par exemple qu'il existe une relation négative entre la taille des personnes et la longueur de leur chevelure. On pourra toujours avancer des arguments plus ou moins psychologiques, mais avant de s'avancer outre mesure, on ferait mieux de revenir sur les conditions du recueil des données et vérifier qu'il n'y a pas d'informations cachées derrière tout cela.

Dans cet exemple, on se rend compte que les hommes et les femmes sont mélangés dans le fichier de données. Or, *en moyenne*, les hommes sont plus grands que les femmes, et inversement, les femmes ont

une chevelure plus longue que les hommes. Le sexe de la personne joue alors le rôle de facteur confondant. L'apparente liaison est un artefact lié à l'existence d'un facteur non maîtrisé.

Dans le cas où le facteur confondant est qualitatif, on détecte facilement le problème en construisant un nuage de points en distinguant les sous-groupes. Étudions plus en détail notre exemple "taille vs. longueur de cheveux" chez les hommes et chez les femmes. Lorsque nous construisons le nuage de points, nous constatons que le nuage des hommes se distingue du nuage des femmes (Figure 2.9). Globalement, une liaison complètement factice apparaît. La corrélation est $\hat{r}_1 = -0.074$ chez les hommes, $\hat{r}_2 = -0.141$ chez les femmes, il passe à $\hat{r} = -0.602$ sur la totalité des individus.



		Cheveux (cm)	Taille (m)
Hommes	1	1.64	1.65
	2	0.32	1.74
	3	1.00	1.76
	4	2.80	1.71
	5	4.35	1.81
	6	2.33	1.71
	7	0.01	1.78
	8	1.75	1.69
	9	3.22	1.77
	10	3.53	1.65
	11	2.55	1.61
	12	3.08	1.77
	13	0.46	1.73
	14	3.22	1.69
	15	2.19	1.75
	16	0.73	1.87
	17	0.16	1.74
	18	0.90	1.69
	19	4.14	1.78
	20	1.61	1.73
Femmes	1	4.66	1.70
	2	3.25	1.54
	3	3.88	1.63
	4	2.84	1.63
	5	4.88	1.44
	6	3.77	1.68
	7	5.64	1.64
	8	4.41	1.63
	9	3.84	1.54
	10	7.58	1.54
	11	7.51	1.62
	12	6.90	1.58
	13	4.76	1.56
	14	6.70	1.50
	15	7.86	1.62

r (hommes)	-0.074
r (femmes)	-0.141
r (global)	-0.602

Fig. 2.9. Nuage de points "taille vs. longueur des cheveux" - Hommes et femmes confondus

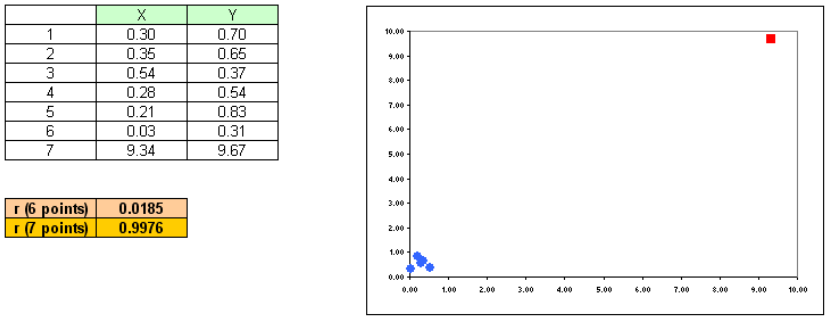
Lorsque le facteur est quantitatif, c'est un peu plus compliqué (exemple : vente de lunettes de soleil et de crèmes glacées, il n'y a pas de lien direct, c'est l'ensoleillement ou la température qui les font varier de manière concomitante). Nous étudierons plus en détail le calcul de la corrélation en contrôlant les effets d'une ou plusieurs tierces variables dans la partie consacrée à la corrélation partielle.

2.6.3 Points aberrants (atypiques)

Dans certains cas, 1 ou 2 points peuvent totalement fausser les résultats. Ces points s'écartent significativement des autres, on parle de points "aberrants" ou "atypiques", dans le sens où ils n'appartiennent (vraisemblablement) pas à la population parente.

Les raisons de l'apparition de ce type d'observations sont multiples : erreur lors du recueil des données (exemple : une personne de 4 ans souscrit à une assurance-vie, en réalité elle a 40 ans) ; un comportement réellement différent (exemple : un sportif tellement dopé qu'il porte les records du monde à des sommets jamais atteints) ; etc.

Le positionnement de ces points par rapport au nuage global laisse croire (ou masque) l'existence d'une liaison manifeste entre les variables. Il existe certes des techniques statistiques destinées à identifier automatiquement les données atypiques, mais force est de constater que des graphiques simples telles que les nuages de points permettent souvent de détecter rapidement les anomalies.



	X	Y
1	0.30	0.70
2	0.35	0.65
3	0.54	0.37
4	0.28	0.54
5	0.21	0.83
6	0.03	0.31
7	9.34	9.67

r (6 points)	0.0185
r (7 points)	0.9976

Fig. 2.10. Influence du point numéro 7 sur le coefficient de corrélation

Dans un premier exemple (Figure 2.10), on note le positionnement totalement atypique de l'individu numéro 7. Si on l'utilise dans les calculs, le coefficient empirique est 0.9976, très proche de liaison linéaire parfaite. Si on le retire c.-à-d. on calcule le coefficient sur les 6 points restants, la corrélation passe à 0.0185. Le point numéro 7 fausse complètement le calcul.

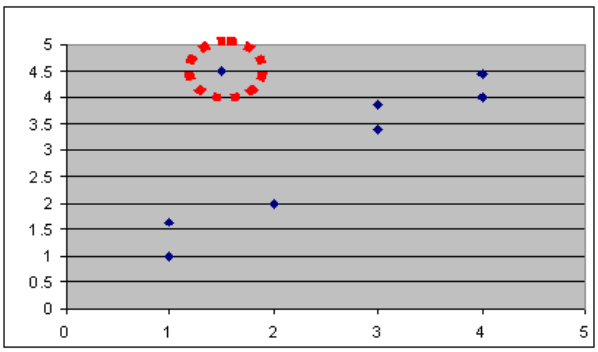


Fig. 2.11. Point aberrant "multivarié"

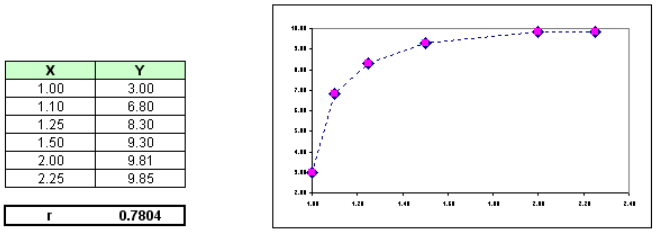
Parfois, le point aberrant est particulièrement sournois. Il est conforme au domaine de définition de X et Y. Mais sur la conjonction (X, Y), il s'écarte du nuage principal (Figure 2.11). Dans cet exemple, le point atypique (entouré de rouge) masque en partie la forte liaison entre X et Y. Les techniques

statistiques de détection univariée des points atypiques [11] sont totalement inopérantes ici. Il faut se tourner vers d'autres procédures. Certaines sont liées à la méthode statistique mise en oeuvre pour analyser les données [12].

2.6.4 Liaison non linéaire

Le coefficient de corrélation sert avant tout à caractériser une liaison linéaire. Lorsqu'elle ne l'est pas, $\hat{r}$ peut nous induire en erreur sur l'existence et l'intensité de la relation entre les variables.

Liaison monotone. Lorsque la liaison est non linéaire mais monotone, le coefficient de corrélation est certes peu adapté mais n'est pas complètement hors de propos : il donne des indications quant à l'existence de la liaison, mais il traduit mal son intensité.



X	Y
1.00	3.00
1.10	6.80
1.25	8.30
1.50	9.30
2.00	9.81
2.25	9.85

r	0.7804

Fig. 2.12. Liaison non linéaire monotone

Dans la figure 2.12, nous constatons visuellement l'existence d'une liaison fonctionnelle quasi parfaite entre X et Y, c'est patent lorsqu'on relie les points. Pourtant le coefficient de corrélation nous annonce $\hat{r} = 0.7804$, indiquant clairement qu'il y a une liaison certes, mais ne rendant pas compte de son intensité. Nous verrons plus loin avec les indicateurs basés sur les rangs comment palier ce problème sans avoir à faire des manipulations compliquées.

Liaison non monotone. Lorsque la liaison est non monotone, c'est la catastrophe : le coefficient de corrélation ne rend compte ni de l'intensité de la liaison, ni même de son existence.

Dans la figure 2.13 (A), on constate immédiatement la forme parabolique de la relation. Pourtant le coefficient de corrélation nous indique $\hat{r}_{xy} = -0.0118$. Effectivement, elle n'est pas linéaire, mais il y a bien une liaison entre X et Y, le coefficient de Pearson est totalement inadapté ici.

Linéarisation par transformation de variables. Une solution évidente, surtout si l'on considère l'exemple précédent, est de proposer une transformation de variables de manière à mettre en exergue une relation linéaire. Dans la figure 2.13 (B), si nous proposons une nouvelle variable $Z = X^2$, la corrélation mesurée en est grandement modifiée $\hat{r}_{zy} = 0.990$. Il y a bien un lien entre les variables, elle est particulièrement forte.

11. Voir `http://tutoriels-data-mining.blogspot.com/2008/05/dtection-univarie-des-points-aberrants.html`

12. Pour la régression multiple, il existe toute une panoplie d'indicateurs assez efficaces - Voir `http://tutoriels-data-mining.blogspot.com/2008/04/points-aberrants-et-influents-dans-la.html`

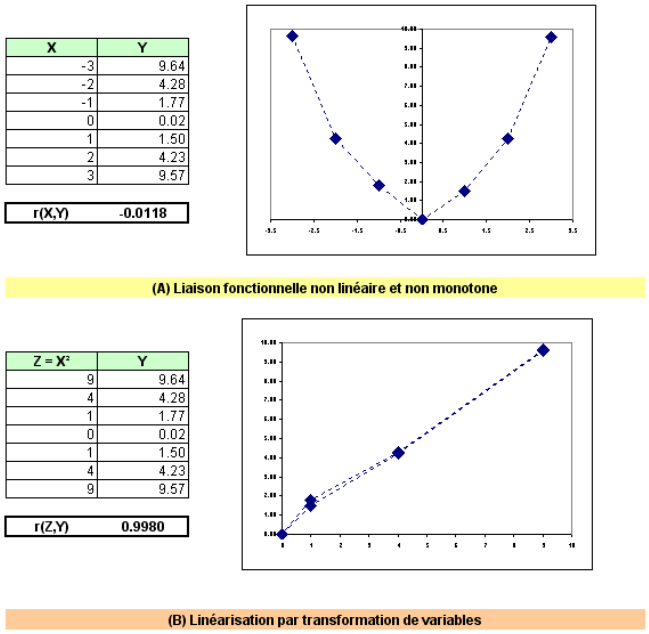


Fig. 2.13. Liaison non linéaire et non monotone

Malheureusement, cette démarche est difficile à reproduire : la fonction de transformation adéquate n'est pas toujours évidente à produire; dans le traitement de gros fichiers où nous avons à manipuler plusieurs dizaines de variables, le nombre de configurations à expertiser est dissuasif.

3

Variations autour de la corrélation

Dans certaines situations, relatives au type des variables, ou consécutives à une transformation des variables, le coefficient de corrélation est simplifié. Son interprétation peut être modifiée et/ou enrichie.

Dans cette partie, nous énumérons quelques unes de ces variantes, les formules et les tests associées. Puis nous montrons leur utilisation et leur interprétation sur un jeu de données.

Quelques références pour cette partie, donnant un positionnement clair des différentes techniques, sont les sites de Garson - `http://www2.chass.ncsu.edu/garson/PA765/correl.htm`, toujours aussi excellents, et de Calkins, de l'Université d'Andrews (USA) - `http://www.andrews.edu/~calkins/math/edrm611/edrm13.htm`

3.1 Corrélation bisériale ponctuelle

3.1.1 Formulation

Le coefficient de corrélation bisériale ponctuelle (*Point biserial correlation coefficient* en anglais [1]) est utilisé pour mesurer la liaison entre une variable dichotomique (X pour fixer les idées) et une variable continue. La variable binaire peut l'être naturellement (ex. sexe = H ou F) ou suite à un découpage en 2 intervalles (ex. revenu, découpé en 2 intervalles). Bien que dans ce second cas, son utilisation ne soit pas très recommandée [2], on préfèrera des indicateurs plus puissants (voir chapitre 3.2).

L'objectif est de mesurer l'association entre Y et X. En calculant le coefficient de Pearson, X étant codé 0/1, nous obtenons exactement le coefficient bisériale ponctuelle. En y regardant de plus près, on se rend compte rapidement qu'il s'agit en réalité de la statistique de la comparaison de moyenne entre 2 échantillons indépendants. On cherche à savoir si dans les sous-groupes définis par X, Y est différent en moyenne.

La corrélation bisériale ponctuelle est définie comme suit pour échantillon de taille n, avec n_1 individus du premier groupe, et n_0 individus du second groupe ($n = n_1 + n_0$)

1. Voir `http://ec.europa.eu/comm/eurostat/research/index.htm?http://www.europa.eu.int/en/comm/eurostat/research/isi/index_fr.htm&1` pour la traduction des termes statistiques
2. Voir `http://en.wikipedia.org/wiki/Point-biserial_correlation_coefficient`

$$r_{pb} = \frac{\bar{y}_1 - \bar{y}_0}{s_{n-1}} \sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{3.1}$$

avec $\bar{y}_1$ et $\bar{y}_0$ les moyennes conditionnelles ; s_{n-1} l'écart type estimé sur l'ensemble de l'échantillon c.-à-d. $s^2_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$.

3.1.2 Test de significativité - 1

En nous basant sur le schéma de la corrélation (section 2.4, nous pouvons tester la significativité du coefficient à l'aide du t_r suivant une loi de Student à $(n_1 + n_0 - 2)$ degrés de liberté

$$t_r = \frac{r_{pb}}{\sqrt{\frac{1-r^2_{pb}}{n_1+n_0-2}}} \tag{3.2}$$

3.1.3 Test de significativité - 2

En nous basant sur le schéma du test de comparaison de moyennes[3] pour échantillons indépendants, nous pouvons vérifier si les moyennes sont significativement différentes dans les sous-groupes. La statistique t_c suit une loi de Student à à $(n_1 + n_0 - 2)$ degrés de liberté

$$t_c = \frac{y_1 - y_0}{s} \tag{3.3}$$

s est l'écart type estimé de l'écart entre les moyennes

- $s^2 = \frac{(n_1-1)s_1^2+(n_0-1)s_0^2}{n_1+n_0-2} (\frac{1}{n_1} + \frac{1}{n_0})$;
- avec $s_j^2 = \frac{1}{n_j-1} \sum_{i=1}^{n_j} (y_i - \bar{y_j})^2$ les variances conditionnelles.

A priori, cette formulation est totalement équivalente à celle basée sur le coefficient de corrélation. Vérifierons cela sur un exemple.

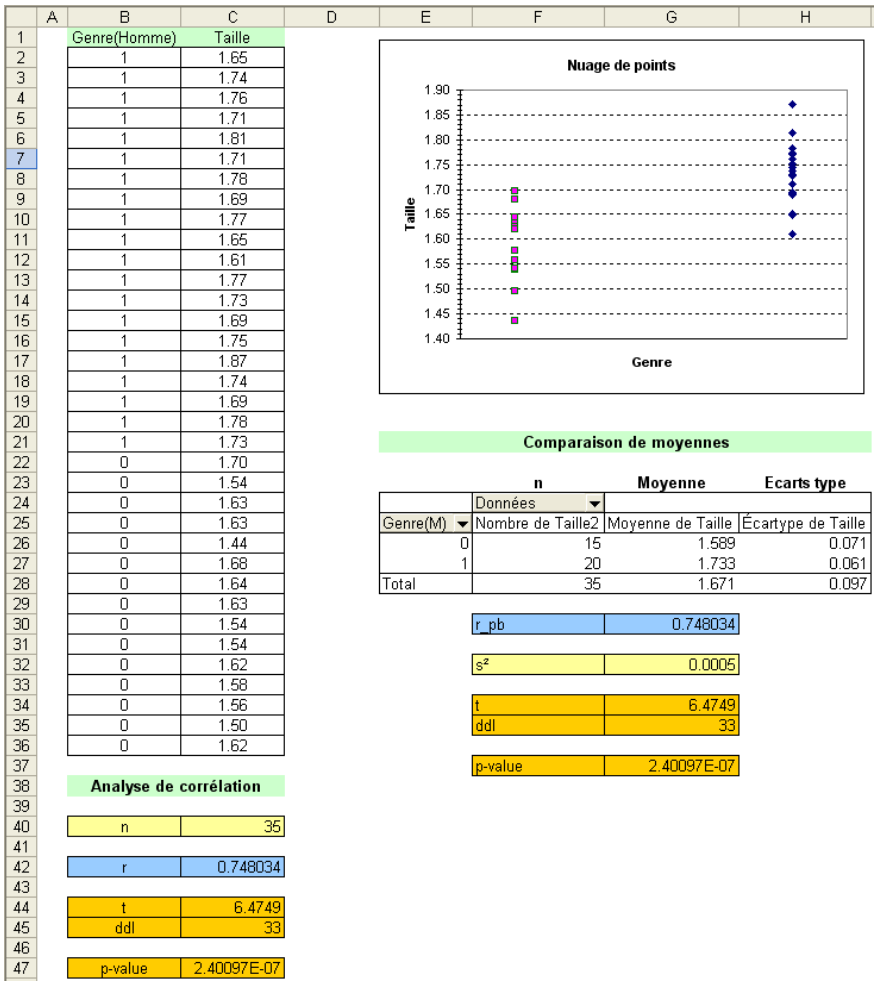
3.1.4 Exemple

Nous voulons vérifier la liaison entre le genre des personnes et leur taille. En d'autres termes nous cherchons à savoir si les hommes, en moyenne, sont plus grands que les femmes. Nous utilisons les données déjà traitées dans la section 2.6.2, nous ne conservons que la taille (Figure 3.1). Nous allons travailler en deux temps, tout d'abord en calculant le coefficient de corrélation sur les données codées, puis en mettant en oeuvre le calcul spécifique sous forme de comparaison de moyennes. Les résultats doivent être cohérents.

Dans les colonnes B et C du tableur, nous avons les données, puis les résultats des calculs basés sur le coefficient de Pearson. Voici les détails des calculs :

- Les hommes sont codés 1, les femmes 0. En soi ça n'a pas d'importance, mais il faudra s'en rappeler lors de l'interprétation du coefficient, le codage détermine le signe du coefficient.

3. `http://en.wikipedia.org/wiki/Student's_t-test`



Genre(Homme)	Taille
1	1.65
1	1.74
1	1.76
1	1.71
1	1.81
1	1.71
1	1.78
1	1.69
1	1.77
1	1.65
1	1.61
1	1.77
1	1.73
1	1.69
1	1.75
1	1.87
1	1.74
1	1.69
1	1.78
1	1.73
0	1.70
0	1.54
0	1.63
0	1.63
0	1.44
0	1.68
0	1.64
0	1.63
0	1.54
0	1.54
0	1.62
0	1.58
0	1.56
0	1.50
0	1.62

Analyse de corrélation

n	35
r	0.748034
t	6.4749
ddl	33
p-value	2.40097E-07

Comparaison de moyennes

	n	Moyenne	Ecarts type
	Données		
Genre(M)	Nombre de Taille2	Moyenne de Taille	Écartype de Taille
0	15	1.589	0.071
1	20	1.733	0.061
Total	35	1.671	0.097

r_pb	0.748034
s²	0.0005
t	6.4749
ddl	33
p-value	2.40097E-07

Fig. 3.1. Corrélation bisériale ponctuelle : taille selon le genre des individus

- Voyons justement le coefficient de Pearson empirique, il est égal à $\hat{r} = 0.748034$. Le signe est positif, cela veut dire qu'en moyenne les hommes sont plus grands que les femmes.
- Le graphique nuage de points confirme cette idée, le nuage des hommes est visuellement plus élevé que celui des femmes, la dispersion étant à peu près la même dans les deux groupes.
- Pour réaliser le test de significativité, nous calculons $t_r = 6.4749$. Il suit une loi de Student à $n - 2 = 33$ degrés de liberté.
- La probabilité critique du test est 2.4×10^{-7}, très petite.
- Au risque 5%, l'hypothèse nulle, il n'y aucun lien entre le genre et la taille, n'est pas compatible avec les données.

Dans les colonnes E, F , G et H du tableur, nous avons les calculs relatifs au coefficient r_{pb} :

- Avec le tableau croisé dynamique, nous avons confirmation des effectifs : $n_0 = 15$ femmes, et $n_1 = 20$ hommes.
- Les moyennes et écarts type dans les sous-groupes sont respectivement ($\bar{y}_0 = 1.589$, $\bar{y}_1 = 1.733$) et ($s_1 = 0.071$, $s_0 = 0.061$).

- Nous en déduisons $s^2 = \frac{19\times 0.071^2+14\times 0.061^2}{20+15-2}(\frac{1}{20}+\frac{1}{15}) = 0.0005$
- Puis $t_c = \frac{1.733-1.589}{\sqrt{0.0005}} = 6.4749$. Nous retrouvons exactement la valeur de t_r
- La distribution et les degrés de liberté étant les mêmes, la p-value du test et la conclusion associée sont identiques.

3.2 Corrélation mutuelle

3.2.1 Formulation et tests

La corrélation mutuelle, que l'on désigne aussi par *corrélation bisériale*[4], est connue sous l'appellation *biserial correlation* en anglais[5]. Elle mesure le lien entre une variable dichotomique X et une variable quantitative Y. La principale différenciation avec la corrélation bisériale ponctuelle est qu'ici, **la variable X doit être issue d'un découpage en 2 intervalles d'une variable continue gaussienne** (voir [2], page 36 ; par exemple : poids bas ou élevé, tension artérielle supérieure à un seuil ou pas, etc.). Attention, dans ce cas le codage de X n'est plus anodin. La valeur 1 correspond naturellement à la fraction élevée (supérieure au seuil de découpage) de la variable sous-jacente.

Remarque 5 (Laquelle privilégier : corrélation bisériale ponctuelle ou corrélation mutuelle ?). La corrélation mutuelle est plus restrictive, si la condition n'est pas respectée, l'inférence statistique est sujette à caution. En revanche, si la condition est remplie, la corrélation mutuelle est plus puissante c.-à-d. elle détectera mieux l'existence d'une relation entre X et Y.

Le coefficient de corrélation mutuelle s'écrit

$$\hat{r}_b = \frac{\bar{y}_1 - \bar{y}_0}{s_{n-1}} \times \frac{n_1 \times n_0}{n^2 \times \lambda_{n_1/n}} \tag{3.4}$$

où

- $s^2_{n-1} = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$ est l'estimation de la variance ;
- $\lambda_{n_1/n}$ est l'ordonnée de la fonction de densité de la loi normale centrée réduite à la coordonnée égale au quantile d'ordre n_1/n (ouf !).

Remarque 6 (Calcul de la quantité $\lambda_{n_1/n}$). Manifestement, mal compris, le calcul de λ est le principal frein à l'utilisation de cet indicateur, qui est très peu présent dans les logiciels. Essayons de détailler la démarche sur un exemple que nous retrouverons dans la section suivante.

- Soit $n_1/n = 23/28 = 0.8214$.
- Nous calculons le quantile d'ordre 0.8214 de la loi normale centrée réduite $u_{0.8214} = 0.9208$.

4. Nous éviterons cette dénomination pour ne pas la confondre avec la corrélation bisériale ponctuelle (ah ces linguistes je vous jure, hein..).

5. `http://ec.europa.eu/comm/eurostat/research/index.htm?http://www.europa.eu.int/en/comm/eurostat/research/isi/index_fr.htm&1`

– Nous appliquons alors la fonction de densité de la loi normale pour obtenir λ c.-à-d.

$$\lambda = f_{\mathcal{N}}(0.9208) = \frac{1}{\sqrt{2\pi}} e^{-\frac{0.9208^2}{2}} = 0.2611$$

Remarque 7 (Violation de l'hypothèse de normalité sous-jacente). Dans certains cas, lorsque la distribution continue sous-jacente de X s'éloigne fortement de la loi normale, bimodale ou très aplatie, $\hat{r}_b$ peut prendre des valeurs supérieures à 1. Ce sont quand même des situations extrêmes. Lorsque la distribution sous-jacente de X est unimodale et raisonnablement symétrique, la procédure est robuste.

Test de significativité. Pour tester la significativité de la corrélation ou calculer les intervalles de confiance, nous pouvons utiliser l'arsenal développé dans les sections 2.4 et 2.5, en substituant la corrélation mutuelle au coefficient de Pearson.

3.2.2 Exemple

Nous cherchons à calculer la corrélation entre la cylindrée dichotomisée ($X = 1$ lorsque cylindrée > 1200, 0 sinon) et la puissance (Y). Dans les études réelles, nous ne disposons que des valeurs binaires de X, nous n'avons pas les valeurs originelles qui ont servi à construire X même si nous savons par ailleurs que la variable sous-jacente est continue.

Détaillons les calculs (Figure 3.2) :

– Nous disposons des effectifs $n = 28$, $n_1 = 23$ et $n_0 = 5$

– A partir du rapport $n_1/n = 0.8214$, nous obtenons le quantile d'ordre 0.8214, soit $u_{0.8214} = 0.9208$. Nous calculons alors l'ordonnée de la fonction de densité de la loi normale centrée réduite à cette coordonnée $f_{\mathcal{N}}(0.9208) = 0.2611$

– Parallèlement à cela, nous calculons l'estimation (non biaisée) de l'écart type $s_{n-1} = 32.2569$, puis les moyennes conditionnelles $m_1 = 87.43$ et $m_0 = 33.00$

– Nous disposons maintenant de tous les éléments pour former la corrélation mutuelle, nous obtenons $\hat{r}_b = 0.9481$

– Le t pour le test de significativité est calculé à l'aide de la formule usuelle $t = \frac{\hat{r}_b}{\sqrt{\frac{1-\hat{r}_b^2}{n-2}}} = 15.2016$

– La corrélation est très hautement significatif, la p-value est très petite. Les données ne sont pas compatibles avec l'hypothèse de nullité du coefficient.

Remarque 8 (Choix de la borne de découpage de la variable continue). Attention, le choix de la borne de découpage (nous avons choisi la valeur 1200 pour cylindrée dans notre exemple) est primordiale. S'il est malheureux, nous pouvons totalement masquer les informations importantes ou, pire, produire des valeurs qui posent problème. Un coefficient de corrélation supérieur à 1 notamment ne manquerait pas de jeter le discrédit sur les techniques que l'on manipule. Il faut donc avoir de bonnes raisons pour effectuer le découpage. Dans la plupart des cas, ce sont les contraintes du domaine ou les exigences de l'étude qui le fixent arbitrairement. Dans notre exemple, on pourrait avancer qu'au delà de la cylindrée 1200, la fiscalité est particulièrement désavantageuse.

	A	B	C	D	E
1	Numero	Modele	Cylindree	x (Cylindrée > 1200)	Puissance (Y)
2	1	Daihatsu Cuore	846	0	32
3	2	Suzuki Swift 1.0 GLS	993	0	39
4	3	Fiat Panda Mambo L	899	0	29
5	4	VW Polo 1.4 60	1390	1	44
6	5	Opel Corsa 1.2i Eco	1195	0	33
7	6	Subaru Vivio 4WD	658	0	32
8	7	Toyota Corolla	1331	1	55
9	8	Opel Astra 1.6i 16V	1597	1	74
10	9	Peugeot 306 XS 108	1761	1	74
11	10	Renault Safrane 2.2. V	2165	1	101
12	11	Seat Ibiza 2.0 GTI	1983	1	85
13	12	VW Golt 2.0 GTI	1984	1	85
14	13	Citroen ZX Volcane	1998	1	89
15	14	Fiat Tempra 1.6 Liberty	1580	1	65
16	15	Fort Escort 1.4i PT	1390	1	54
17	16	Honda Civic Joker 1.4	1396	1	66
18	17	Volvo 850 2.5	2435	1	106
19	18	Ford Fiesta 1.2 Zetec	1242	1	55
20	19	Hyundai Sonata 3000	2972	1	107
21	20	Lancia K 3.0 LS	2958	1	150
22	21	Mazda Hachtback V	2497	1	122
23	22	Mitsubishi Galant	1998	1	66
24	23	Opel Omega 2.5i V6	2496	1	125
25	24	Peugeot 806 2.0	1998	1	89
26	25	Nissan Primera 2.0	1997	1	92
27	26	Seat Alhambra 2.0	1984	1	85
28	27	Toyota Previa salon	2438	1	97
29	28	Volvo 960 Kombi aut	2473	1	125
30					
31					
32	n	28			
33	n_1	23			
34	n_0	5			
35					
36	n_1/n	0.8214			
37	u(n_1/n)	0.9208			
38	lambda	0.2611			
39					
40	s_{n-1}	32.2569			
41					
42	m_1	87.43			
43	m_0	33.00			
44					
45	**Coefficient de corrélation mutuelle (bisériale)**				
46	r_b	0.9481			
47					
48	t	15.2016			
49	p-value	1.88485E-14			

Coefficient de Pearson (sur les variables originelles)

r	0.9475

Coefficient bisériale ponctuelle

r_pb	0.6582

r_b (vérification)	0.9481

Fig. 3.2. Corrélation mutuelle : cylindrée vs. puissance

3.2.3 Commentaires sur la puissance de $\hat{r}_b$ par rapport $\hat{r}_{pb}$

Par rapport à la corrélation bisériale ponctuelle, la corrélation mutuelle tient compte explicitement du fait que la variable sous-jacente à X est continue et gaussienne. Ce surcroît d'information utilisé dans les calculs la rend particulièrement puissante lorsque l'assertion est vraie. Dans la pratique, on se rend compte qu'il y a une formule de passage entre les 2 indicateurs ([2], page 37)

$$\hat{r}_b = \hat{r}_{pb}\sqrt{\frac{n_1 n_0 (n-1)}{\lambda^2_{n_1/n} \times n^3}} \tag{3.5}$$

Nous avons effectué plusieurs vérifications pour notre exemple précédent (Figure 3.2). Détaillons les résultats :

- En calculant le coefficient de Pearson sur les données originelles (la variable X non dichotomisée), nous obtenons $\hat{r} = 0.9475$. Rappelons que la corrélation mutuelle est $\hat{r}_b = 0.9481$. Il est quand même remarquable que cette dernière puisse reconstituer avec une telle précision les résultats en se basant sur la variable dichotomisée et une hypothèse de normalité de la variable sous-jacente.
- La corrélation bisériale ponctuelle, basée uniquement sur la variable dichotomisée, qu'importe qu'elle soit intrinsèquement qualitative ou non, sous-estime fortement l'intensité du lien. En effet, on obtient $\hat{r}_{pb} = 0.6582$. Même si elle reste significative, elle est loin de traduire la liaison réelle qui existe entre les variables cylindrée et puissance, évidente lorsque l'on construit le graphique nuage de points associé (Figure 2.5).
- En appliquant la formule de passage ci-dessus (équation 3.5), nous retrouvons exactement la valeur de la corrélation mutuelle [la case $\hat{r}_b$(vérification)].

Concernant le passage entre la corrélation mutuelle et la corrélation bisériale ponctuelle, on montre que

$$\sqrt{\frac{n_1 n_0 (n-1)}{\lambda^2_{n_1/n} n^3}} \geq 1.25$$

La corrélation mutuelle est toujours supérieure à la corrélation bisériale ponctuelle ($\hat{r}_b > \hat{r}_{pb}$). Elle a tendance à mieux mettre en évidence les écarts à l'hypothèse nulle. Cela n'est pas sans dangers, comme nous le signalions plus haut, dans certaines situations $\hat{r}_b$ peut prendre des valeurs supérieures à 1.

3.3 Le coefficient ϕ

3.3.1 Formulation et tests

Le coefficient ϕ est utilisé pour mesurer le degré de liaison entre 2 variables binaires codées 0/1. Les variables peuvent être dichotomiques par nature (sexe = H/F) ou dichotomisées (découpage en 2 intervalles d'une variable continue). Dans ce dernier cas, il est moins puissant, on préférera se tourner vers la corrélation tetrachorique (section 3.3.3).

Calcul basé sur le coefficient de Pearson. Une première manière très simple de calculer le coefficient ϕ est de calculer le coefficient de Pearson sur les variables codées 0/1. Aucune correction n'est nécessaire, nous obtenons directement la valeur adéquate.

Calcul basé sur le tableau de contingence. Comme les variables sont censées être dichotomiques qualitatives c.-à-d. les modalités ne sont pas ordonnées. Nous pouvons élaborer un tableau de contingence

croisant les modalités de X et Y. Et calculer l'indicateur ϕ dessus. Nous nous rapprochons en cela des mesures d'association entre variables qualitatives [6]

Partons du tableau de contingence générique 2×2 pour établir les formules (Tableau 3.1). En ligne les modalités de Y, en colonne celles de X.

Y vs. X	1	0
1	a	b
0	c	d

Tableau 3.1. Tableau générique 2×2

Le coefficient ϕ s'écrit :

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{3.6}$$

Le codage 0 ou 1 détermine le signe de ϕ, il n'a pas d'incidence sur la valeur absolue du coefficient. Cela permet de détecter les attraction ou les répulsions entre les modalités.

Test de significativité. Pour tester la significativité de la corrélation ou calculer les intervalles de confiance, nous pouvons utiliser l'arsenal développé dans les sections 2.4 et 2.5, en substituant le coefficient ϕ au coefficient de Pearson.

3.3.2 Exemple

Reprenons notre exemple de la puissance et de la cylindrée (Figure 2.5). Les deux variables ont été maintenant dichotomisées, nous avons choisi le seuil 1800 pour la variable cylindrée, 75 pour "puissance". Ce faisant nous perdons de l'information car ϕ ne tient pas compte de la nature continue des variables sous-jacentes. Nous essaierons de voir justement dans quelle mesure la perte d'information est préjudiciable.

Détaillons notre feuille de calcul (Figure 3.3) :

- Dans les colonnes C et D, nous avons les variables originales. En E et F, les variables dichotomisées.
- Dans la partie droite, sous le tableau de données, nous avons classiquement calculé le coefficient de Pearson sur données dichotomiques.Nous obtenons $\hat{r} = 0.9309$. Le test de significativité propose $t = 13.0$. L'hypothèse nulle d'absence de liaison n'est pas compatible avec les données.
- Voyons maintenant la partie gauche. Nous avons formé le tableau de contingence, puis à partir de la formule 3.6, nous avons obtenu $\phi = 0.9309$. La valeur coïncide avec le coefficient précédent. C'est heureux.

Rappelons que la corrélation sur les variables continues originelles est $\hat{r}_{cyl,puiss} = 0.9475$. Après découpage en 2 intervalles des variables, nous retrouvons quand même l'intensité de la liaison avec $\hat{r} = 0.9309$.

6. Rakotomalala, R., *Etude des dépendances - Variables qualitatives*, `http://eric.univ-lyon2.fr/~ricco/cours/cours/Dependance_Variables_Qualitatives.pdf`. Voir la section 4.1 concernant le coefficient ϕ et sa relation avec le coefficient de corrélation.

	A	B	C	D	E	F	G
1					(Cylindrée > 1800)	(Puissance > 75)	
2	Numero	Modele	Cylindree	Puissance	X	Y	
3	1	Daihatsu Cuore	846	32	0	0	
4	2	Suzuki Swift 1.0 GLS	993	39	0	0	
5	3	Fiat Panda Mambo L	899	29	0	0	
6	4	VW Polo 1.4 60	1390	44	0	0	
7	5	Opel Corsa 1.2i Eco	1195	33	0	0	
8	6	Subaru Vivio 4WD	658	32	0	0	
9	7	Toyota Corolla	1331	55	0	0	
10	8	Opel Astra 1.6i 16V	1597	74	0	0	
11	9	Peugeot 306 XS 108	1761	74	0	0	
12	10	Renault Safrane 2.2. V	2165	101	1	1	
13	11	Seat Ibiza 2.0 GTI	1983	85	1	1	
14	12	VW Golt 2.0 GTI	1984	85	1	1	
15	13	Citroen ZX Volcane	1998	89	1	1	
16	14	Fiat Tempra 1.6 Liberty	1580	65	0	0	
17	15	Fort Escort 1.4i PT	1390	54	0	0	
18	16	Honda Civic Joker 1.4	1396	66	0	0	
19	17	Volvo 850 2.5	2435	106	1	1	
20	18	Ford Fiesta 1.2 Zetec	1242	55	0	0	
21	19	Hyundai Sonata 3000	2972	107	1	1	
22	20	Lancia K 3.0 LS	2958	150	1	1	
23	21	Mazda Hachtback V	2497	122	1	1	
24	22	Mitsubishi Galant	1998	66	1	0	
25	23	Opel Omega 2.5i V6	2496	125	1	1	
26	24	Peugeot 806 2.0	1998	89	1	1	
27	25	Nissan Primera 2.0	1997	92	1	1	
28	26	Seat Alhambra 2.0	1984	85	1	1	
29	27	Toyota Previa salon	2438	97	1	1	
30	28	Volvo 960 Kombi aut	2473	125	1	1	
31							
32							
33			**Coefficient phi**				
34							

	A	B	C	D
35	Nombre de Y	X		
36	Y	0	1	Total
37	0	13	1	14
38	1		14	14
39	Total	13	15	28
40				
41		Phi	0.9309	
42				

	n	28
	Corr.Pearson	0.9309
	t	13.0000
	ddl	26
	p-value	6.96946E-13

Fig. 3.3. Corrélation ϕ : cylindrée vs. puissance dichotomisées

Dans ce cas il y a peu de pertes d'informations. Ce n'est pas étonnant, les seuils ont été judicieusement choisis, ils se rapprochent, à peu près, du barycentre du nuage de points (Figure 2.5). Si nous avions choisi des seuils qui ne sont pas en correspondance, par exemple 900 pour la cylindrée et 100 pour la puissance, nous aurions obtenu $\hat{r} = 0.3523$, laissant à croire que le lien est faible. Ce qui est totalement erroné bien sûr.

Remarque 9 (Découper en intervalles peut même être profitable). Encore une fois, la préparation des données, en l'occurrence le choix des bornes lorsque l'on découpe les données, est donc très important pour ce type d'indicateur. Il faut faire très attention. Mais a contrario, un choix judicieux des bornes peut être profitable à l'analyse. Si la relation est fortement non linéaire, le coefficient de Pearson sur les variables originelles est faussé. Le découpage en intervalles peut aider à mieux mettre en évidence l'existence de la liaison.

3.3.3 Corrélation tetrachorique

Lorsque les deux variables ont été dichotomisées à partir d'un couple de variables distribuées selon une loi normale bivariée, on privilégiera le coefficient tetrachorique qui est plus puissant (*Tetrachoric coefficient* en anglais [7]).

Ce coefficient s'appuie sur l'hypothèse de normalité sous jacente pour corriger le coefficient ϕ (équation 3.6). *Grosso modo*, le numérateur reste le même, le dénominateur doit tenir compte en revanche de la distribution normale en intégrant de nouveau l'ordonnée de la loi normale centrée et réduite pour les quantiles des proportions $\frac{a+b}{n}$ et $\frac{a+c}{n}$. Le calcul est loin d'être trivial cependant [8], on peut avoir des problèmes lorsque l'on s'éloigne trop de l'hypothèse de normalité. Ce coefficient est très peu utilisé dans la pratique.

3.4 ρ de Spearman

Fondamentalement, le coefficient de Spearman est aussi un cas particulier du coefficient de Pearson, calculé à partir des transformations des variables originelles. Mais il présente l'avantage d'être **non paramétrique**. L'inférence statistique ne repose plus sur la normalité bivariée du couple de variables (X, Y). Nous pouvons bien entendu mettre en oeuvre tous les tests mis en avant dans la section 2.5, y compris ceux relatifs à la comparaison de coefficients.

3.4.1 Principe

L'idée est de substituer aux valeurs observées leurs rangs. Nous créons donc deux nouvelles colonnes dans notre tableau : $R_i = Rang(x_i)$, correspond au rang [9] de l'observation x_i dans la colonne des X ; et $S_i = Rang(Y_i)$.

Le ρ de Spearman est ni plus ni moins que le coefficient de Pearson calculé sur les rangs.

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}} \tag{3.7}$$

Compte tenu de certaines propriétés des rangs (par ex. $\bar{S} = \bar{R} = \frac{n+1}{2}$; voir [3], pages 105 à 108), nous pouvons déduire une expression simplifiée

$$\hat{\rho} = \frac{12\sum_{i=1}^{n} R_i S_i}{n(n^2-1)} - \frac{3(n+1)}{n-1} \tag{3.8}$$

7. `http://ec.europa.eu/comm/eurostat/research/index.htm?http://www.europa.eu.int/en/comm/eurostat/research/isi/index_fr.htm&1`

8. Voir `http://ourworld.compuserve.com/homepages/jsuebersax/tetra.htm` concernant les fondements et les interprétations de la mesure ; voir `http://lib.stat.cmu.edu/apstat/116` sur son mode de calcul dans les logiciels de statistique

9. La plus petite valeur prend le rang 1, la plus grande le rang n

Enfin, si nous définissons D_i telle que $D_i = R_i - S_i$ est l'écart entre les rangs, nous obtenons une autre expression équivalente

$$\hat{\rho} = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} \tag{3.9}$$

Attention, pour ces équations simplifiées, il est nécessaire d'**introduire une correction lorsqu'il y a des ex-aequo dans les données**, surtout s'ils sont assez nombreux. Nous reviendrons en détail sur les corrections à introduire plus loin (section 3.4.5).

Le ρ de Spearman est une variante du coefficient de Pearson, il en reprend les propriétés essentielles, à savoir : $-1 \leq \rho \leq +1$; il prend la valeur 0 lorsque les variables sont indépendantes.

3.4.2 Un exemple

	A	B	C	D	E	F	G	H
1	Numéro	Taille (m)	Poids (kg)	R_i	S_i	R_i x S_i	D_i	D_i^2
2	1	1.697	77.564	15	14	210	1	1
3	2	1.539	55.000	3	1	3	2	4
4	3	1.629	76.657	10	12	120	-2	4
5	4	1.633	62.596	11	6	66	5	25
6	5	1.500	58.068	2	3	6	-1	1
7	6	1.679	72.575	14	11	154	3	9
8	7	1.643	82.000	13	15	195	-2	4
9	8	1.626	76.667	9	13	117	-4	16
10	9	1.543	58.060	5	2	10	3	9
11	10	1.542	71.668	4	10	40	-6	36
12	11	1.621	68.039	8	8	64	0	0
13	12	1.577	70.060	7	9	63	-2	4
14	13	1.557	61.689	6	5	30	1	1
15	14	1.496	67.585	1	7	7	-6	36
16	15	1.637	59.874	12	4	48	8	64
17					Somme	**1133**	Somme	**214**
18	n	15						
19								
20	r Pearson	0.58452						
21								
22								
23	rho (1)	0.61786						
24	rho (2)	0.61786						
25	rho (3)	0.61786						
26								
27	t	2.83320						
28	p-value	0.01410						
29								
30	U	2.31181						
31	p-value	0.02079						
32								
33								

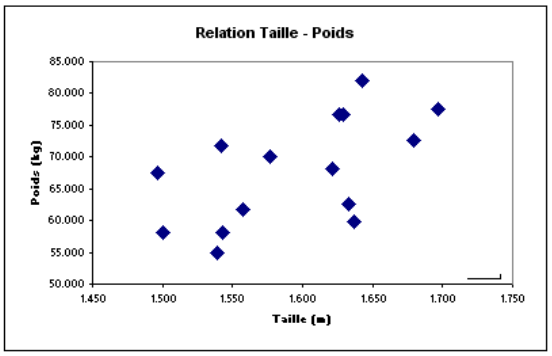


Fig. 3.4. Calcul du ρ de Spearman sur une relation "taille - poids"

Nous reprenons notre exemple du lien entre la taille et le poids. Nous avons modifié les données de manière à éviter les ex-aequo :

- Nous avons tout d'abord formé le nuage de points. Il semble y avoir une liaison entre les 2 variables.
- Le coefficient de corrélation de Pearson est de $\hat{r} = 0.58452$.
- Dans la colonne D et E, nous calculons respectivement les rangs R_i et S_i

- Nous calculons alors le ρ avec la formule 3.7 c.-à-d. en appliquant directement la formule de Pearson sur les rangs. Nous obtenons $\hat{\rho} = 0.61786$
- Dans la colonne F, nous formons le produit $R_i \times S_i$, nous obtenons la somme $\sum_i R_i S_i = 1133$. A partir de la formule 3.8, nous produisons $\hat{\rho} = 0.61786$. La même valeur que précédemment.
- Enfin, en colonne G, nous calculons l'écart D_i et nous formons la colonne D_i^2. La somme $\sum_i D_i^2 = 214$. En appliquant la formule 3.9, la troisième estimation $\hat{\rho} = 0.61786$ est totalement cohérente avec les précédentes.

3.4.3 Distribution et tests

Nous pouvons utiliser la transformation de Fisher pour calculer les intervalles de confiance et réaliser les tests de comparaison.

Concernant le test de significativité, nous nous appuyons sur le t de Student lorsque n est de l'ordre de 20 à 30

$$t = \frac{\hat{\rho}}{\sqrt{\frac{1-\hat{\rho}^2}{n-2}}}$$

Ou même utiliser une approximation normale, plus simple, lorsque $n > 35$

$$U = \frac{\hat{\rho}}{\sqrt{\frac{1}{n-1}}} = \hat{\rho} \times \sqrt{n-1}$$

Remarque 10 (Approximation asymptotique). Les valeurs de n ci-dessus correspondent avant tout à un ordre d'idée. Les ouvrages divergent à ce sujet, Dodge et Rousson rapportent que l'approximation normale suffit dès que $n > 10$ (voir [3], page 107) ; Siegel et Castellan, eux, rapportent qu'on peut s'appuyer sur l'approximation normale lorsque n est autour de 20 - 25 (voir [9], page 243). Ce qui est sûr, c'est que lorsque les effectifs sont vraiment faibles ($4 \leq n \leq 10$), nous avons intérêt à utiliser des tables spécifiques pour les tests de significativité (voir la table 24 dans [1] ; la table Q dans [9] ; ou `http://www.sussex.ac.uk/Users/grahamh/RM1web/Rhotable.htm`).

Exemple numérique. Nous avons mis en oeuvre les deux approximations dans notre exemple ci-dessus (Figure 3.4). Nous avons $t = 2.83320$ avec une p-value de 0.01410 pour le premier test ; $U = 2.31181$ avec p-value = 0.02079 pour le second. Les résultats ne sont guères différents au final, ils aboutissent à la même conclusion, le rejet de l'hypothèse de nullité du coefficient au risque 5%.

3.4.4 D'autres propriétés réjouissantes

Dans la pratique, on se rend compte que le ρ de Spearman cumule les bonnes qualités. Il devrait être privilégié dès que l'on effectue des traitement automatisés. Il évite bien des écueils qui faussent souvent les valeurs produites par le coefficient de Pearson.

Test non paramétrique. Il est non paramétrique, il n'est donc pas nécessaire de faire des hypothèses sur les distributions sous-jacentes de X et Y. Mais lorsque le couple (X, Y) est distribué selon une loi normale bivariée, il est quasiment aussi puissant que le coefficient de Pearson. Les deux indicateurs proposent des valeurs similaires, il est dès lors possible d'interpréter le carré du coefficient de Spearman en termes de variance expliquée.

Traitement des données ordinales. Toujours conséquence du fait qu'il soit non paramétrique, le ρ de Spearman peut traiter les variables intrinsèquement ordinales : un indice de satisfaction, une appréciation ou une note attribuée, etc. L'inférence statistique (tests, intervalles de confiance) n'est pas modifiée.

Liaison monotone non linéaire. Très intéressant dans la pratique, le ρ de Spearman peut caractériser d'une liaison non-linéaire monotone, à la différence du coefficient de Pearson qui ne retranscrit que les relations linéaires. Cela nous évite d'avoir à effectuer le choix douloureux de la fonction de transformation lors de la tentative de linéarisation de l'association. La transformation par les rangs est suffisamment générique pour que l'on puisse rendre compte de l'existence d'une liaison monotone.

De manière générale, une forte disparité entre $\hat{\rho}$ et $\hat{r}$ devrait nous alerter quant à la non linéarité de la relation entre X et Y.

X	Y	R	S
1.00	3.00	1.00	1.00
1.10	6.80	2.00	2.00
1.25	8.30	3.00	3.00
1.50	9.30	4.00	4.00
2.00	9.81	5.00	5.00
2.25	9.85	6.00	6.00

r	0.7804	rho	1.0000

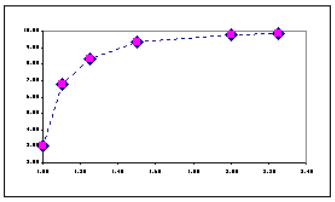


Fig. 3.5. Avantage du ρ de Spearman sur une relation non linéaire monotone

Reprenons l'exemple illustratif de la section 2.6 (Problèmes et cas pathologiques). Rappelons nous, malgré une liaison visuellement évidente, le coefficient de Pearson nous annonçait une corrélation $\hat{r} = 0.7804$. Nous avons remplacé les valeurs initiales par les rangs, puis nous avons calculé le coefficient de Spearman, la liaison parfaite est maintenant bien détectée (Figure 3.5). Ceci s'explique en partie par le fait que **le passage aux rangs symétrise les distributions**. En effet, dans notre exemple, la distribution initiale de la variable en ordonnée est très asymétrique, faussant le coefficient de Pearson.

Le ρ de Spearman a quand même des limites. Lorsque la liaison est non monotone, il n'est pas opérant. Il faut se tourner vers une transformation de variable spécifique inspirée par le graphique nuage de points ou utiliser un indicateur adapté tel que le rapport de corrélation (section 3.6).

Robustesse face aux points atypiques.

Autre caractéristique très intéressante du coefficient de Spearman, sa robustesse face aux points aberrants, même lorsque l'effectif est faible.

	X	Y	R	S
1	0.30	0.70	4	5
2	0.35	0.65	5	4
3	0.54	0.37	6	2
4	0.28	0.54	3	3
5	0.21	0.83	2	6
6	0.03	0.31	1	1
7	9.34	9.67	7	7

r (6 points)	0.0185
r (7 points)	0.9976

rho	0.39286

t	0.95526
p-value	0.38332

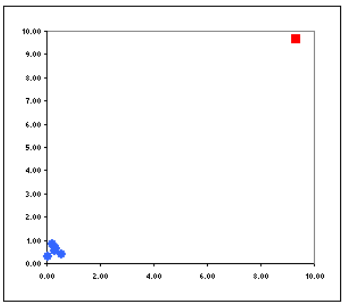


Fig. 3.6. Avantage du ρ de Spearman concernant les points atypiques

Reprenons l'exemple présenté plus haut (section 2.6, figure 2.10). Nous avions noté que le coefficient de Pearson pouvait être fortement affecté par l'existence d'un point extrême. Nous avons transformé les données en rangs, ce faisant nous avons lissé les écarts entre les valeurs. Nous calculons ρ sur l'ensemble des observations, nous obtenons $\hat{\rho} = 0.39286$, et nous notons surtout que le coefficient n'est pas significativement différent de 0, avec $t = 0.95526$ et une p-value $= 0.38332$.

3.4.5 Traitement des ex aequo

Lorsqu'il y a beaucoup d'ex-aequo dans les données, nous affectons les rangs moyens aux observations portant des valeurs identiques. Il faut alors ajuster le coefficient de Spearman lorsque nous voulons utiliser l'équation 3.9 (voir [9], pages 239 à 241). La correction est d'autant plus sensible que le nombre de valeurs identiques est élevé pour X et Y.

Dans ce qui suit, nous explicitons le processus pour la variable X. Les calculs sont exactement les mêmes pour la variable Y.

Rangs moyens. Lors de la transformation des données en rangs, nous devons tenir compte maintenant des ex-aequo. Pour un échantillon de taille n, admettons qu'il n'y ait que G valeurs différentes. Remarquons que si $G = n$, cela veut dire qu'il n'y pas d'ex aequo dans nos données.

Au départ nous affectons les rangs aux observations selon la procédure habituelle. Dans un deuxième temps, nous effectuons un nouveau passage sur les données, nous attribuons aux individus portant des valeurs identiques la moyenne des rangs associés.

Prenons un petit exemple pour détailler cela (Figure 3.7). Nous avons 12 observations triés selon la valeur de X. Nous attribuons le rang normalement (Rangs bruts) en utilisant la fonction RANG(...) d'EXCEL. Nous notons que plusieurs observations ont des valeurs identiques (A,B), (D,E,F) et (J,K) [10]. Nous effectuons un second passage sur les données, nous calculons et attribuons la moyenne de leur rangs

10. La procédure est totalement générique bien sûr, nous pouvons avoir 10 valeurs identiques

Individu	X	Rangs bruts	Rangs moyens
A	0	1	1.5
B	0	2	1.5
C	1	3	3
D	2	4	5
E	2	5	5
F	2	6	5
G	5	7	7
H	6	8	8
I	7	9	9
J	8	10	10.5
K	8	11	10.5
L	12	12	12

Fig. 3.7. Calcul des rangs moyens

aux individus portant les mêmes valeurs. Ici, A et B ont la même valeur, ils portent respectivement les rangs 1 et 2, nous leur affectons au final le rang moyen $\frac{1+2}{2} = 1.5$. Pour D, E et F nous effectuons le calcul $\frac{4+5+6}{3} = 5$. Et pour J et K, nous calculons $\frac{10+11}{2} = 10.5$.

Facteur de correction. Pour calculer le facteur de correction T_x, nous recensons les G valeurs distinctes parmi les rangs moyens, pour chaque valeur nous comptons son nombre d'apparition t_g. Nous produisons alors la quantité T_x qui sera introduite dans la formule du coefficient de Spearman (il en sera de même pour T_y, facteur de correction pour Y)

$$T_x = \sum_{g=1}^{G} (t_g^3 - t_g) \tag{3.10}$$

Reprenons notre exemple ci-dessus (Figure 3.7). Nous avons $n = 12$ et $G = 8$. Pour chaque valeur du rang moyen, nous associons le nombre d'occurrence t_g. Nous appliquons la formule 3.10 pour obtenir $T_x = 36$ (Figure 3.8).

	G	8
Valeur (Rang.Moyen)	**t_g**	**Calc (Tx)**
1.5	2	6
3	1	0
5	3	24
7	1	0
8	1	0
9	1	0
10.5	2	6
12	1	0
	Somme = Tx	**36**

Fig. 3.8. Calcul du facteur de correction pour le ρ de Spearman

Coefficient de Spearman corrigé. Enfin, il nous faut introduire le facteur de correction dans le calcul du ρ de Spearman (Equation 3.9) (voir [9], page 239, équation 9.7)

$$\hat{\rho} = \frac{(n^3 - n) - 6\sum_{i=1}^{n} d_i^2 - (T_x + T_y)/2}{\sqrt{(n^3 - n)^2 - (T_x + T_y)(n^3 - n) + T_x T_y}} \tag{3.11}$$

Remarquons que s'il n'y a pas d'ex-aequo en X et en Y, nous aurons $T_x = T_y = 0$, la formule 3.11 sera totalement équivalente (après quelques simplifications) à la formule 3.9.

Complétons notre exemple avec les valeurs de Y. Pour rendre l'exposé plus clair, il n'y a pas d'ex aequo sur cette seconde variable, de facto $T_y = 0$ (Figure 3.8). Nous construisons les rangs S_i, nous calculons les écarts $D_i = R_i - S_i$. Reste à produire D_i^2 que nous introduisons dans l'équation 3.11 :

$$\hat{\rho} = \frac{(12^3 - 12) - 6 \times 129 - (36 + 0)/2}{\sqrt{(12^3 - 12)^2 - (36 + 0)(12^3 - 12) + 36 \times 0}} = 0.5442$$

Individu	X	R	Y	S	D_i	D_i^2
A	0	1.5	42	3	-1.5	2.25
B	0	1.5	46	4	-2.5	6.25
C	1	3	39	2	1	1
D	2	5	37	1	4	16
E	2	5	65	8	-3	9
F	2	5	88	11	-6	36
G	5	7	86	10	-3	9
H	6	8	56	6	2	4
I	7	9	62	7	2	4
J	8	10.5	92	12	-1.5	2.25
K	8	10.5	54	5	5.5	30.25
L	12	12	81	9	3	9
					Somme	**129**

Fig. 3.9. Tableau de calcul du ρ de Spearman lorsqu'il y a des ex-aequo

Remarque 11 (Traitement des ex-aequo pour le coefficient de Pearson sur les rangs). Comme nous le signalions plus haut, il est possible d'obtenir le ρ de Spearman en calculant le r de Pearson sur les rangs. Avec cette stratégie, lorsqu'il y a des ex aequo dans les données, nous utilisons toujours le principe des rangs moyens. En revanche il n'est pas nécessaire de corriger le coefficient obtenu [11]. Dans notre exemple ci-dessus (Figure 3.9), si nous appliquons la formule de la corrélation empirique (Equation 2.8) sur les colonnes des rangs moyens R et S, nous obtenons directement la bonne valeur de $\hat{\rho} = 0.5442$.

3.5 τ de Kendall

Le τ de Kendall n'est pas à proprement parler une variante du coefficient de Pearson. On n'applique pas la formule sur des données recodées. Il repose sur un principe très différent, il s'interprète également de manière différente. Nous le présentons dans ce support car il est très largement diffusé, et certains auteurs s'accordent à dire qu'il est meilleur que le ρ de Spearman [12]. Nous ne rentrerons pas dans cette polémique. En revanche, nous ne pouvons pas passer à côté de cette mesure, d'autant plus qu'elle est aussi non paramétrique.

11. `http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient`
12. Voir par exemple `http://www.rsscse.org.uk/ts/bts/noether/text.html` ; voir aussi [6], page 332

3.5.1 Principe et interprétation

Le τ de Kendall est défini pour mesurer l'association entre variables ordinales, typiquement des classement (ou rangs) affectés par des juges. Son champ d'application couvre donc parfaitement celui du ρ de Spearman.

Le coefficient de Kendall repose sur la notion de paires discordantes et concordantes [13] :

1. On dit que les paires observations i et j sont concordantes si et seulement si ($x_i > x_j$ alors $y_i > y_j$) **ou** ($x_i < x_j$ alors $y_i < y_j$). Nous pouvons simplifier l'écriture avec $(x_i - x_j) \times (y_i - y_j) > 0$
2. On dit que les paires sont discordantes lorsque ($x_i > x_j$ alors $y_i < y_j$) **ou** ($x_i < x_j$ alors $y_i > y_j$), en d'autres termes $(x_i - x_j) \times (y_i - y_j) < 0$

Pour un échantillon de taille n, soit P (resp. Q) le nombre de paires concordantes (resp. discordantes). Le τ de Kendall est défini de la manière suivante

$$\hat{\tau} = \frac{P - Q}{\frac{1}{2}n(n-1)} \tag{3.12}$$

Le dénominateur représente le nombre total de paires possibles c.-à-d.

$$\frac{1}{2}n(n-1) = \binom{n}{2}$$

Remarque 12 (Données continues, données ordinales). Notons qu'il est possible de calculer directement $\hat{\tau}$ sur des données continues (X et Y) sans qu'il soit nécessaire de les transformer en rangs. Le τ de Kendall s'applique naturellement aussi lorsqu'une des variables est continue, l'autre ordinale.

Interprétation. Le τ de Kendall s'interprète comme le degré de correspondance entre 2 classements (ou 2 notations). Si toutes les paires sont concordantes c.à-d. le classement selon X concorde systématiquement avec le classement selon Y, $\tau = 1$; si toutes les paires sont discordantes, $\tau = -1$; enfin, si les deux classements sont totalement indépendants, $\tau = 0$.

Surtout, et c'est sa principale différenciation avec le ρ de Spearman, le τ de Kendall se lit comme une probabilité. Il est le fruit de la différence entre 2 probabilités : celle d'avoir des paires concordantes et celle d'avoir des paires discordantes. Ainsi, lorsque $\tau = 0$, une paire d'observations a autant de chances d'être concordante que d'être discordante.

Le τ de Kendall théorique, calculé sur la population, est défini par (voir [7], 138)

$$\tau = 2 \times P[(x_i - x_j) \times (y_i - y_j) > 0] - 1 \tag{3.13}$$

Calcul pratique. La manière la plus simple de calculer $\hat{\tau}$ est de trier les données selon X, puis de comptabiliser la quantité suivante

13. http://en.wikipedia.org/wiki/Concordant_pairs

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \nu_{ij}$$

où

$$\nu_{ij} = \begin{cases} +1 \text{ , si} & y_i < y_j \\ -1 \text{ , si} & y_i > y_j \end{cases} \tag{3.14}$$

et

$$\nu_i = \sum_{j=i+1}^{n} \nu_{ij}$$

ν_i est l'écart entre le nombre de paires concordantes et discordantes relativement à l'observation i.

S est donc l'écart entre le nombre total de paires concordantes, et le nombre total de paires discordantes c.-à-d. $S = P - Q$. Nous pouvons dès lors ré-écrire le coefficient de Kendall

$$\hat{\tau} = \frac{S}{\frac{1}{2}n(n-1)} = \frac{2S}{n(n-1)} \tag{3.15}$$

Un exemple. Détaillons les calculs sur exemple. Nous limitons les effectifs à $n = 6$ car les calculs deviennent rapidement inextricables. Nous mettons en relation la taille et le poids des 6 plus petits individus du fichier (Figure 3.4). Les données sont triées selon la taille, nous allons calculer les quantités ν_{ij}, ν_i et S (Figure 3.10).

Numéro	Taille (m)	Poids (kg)	n°1	n°2	n°3	n°4	n°5
1	1.496	67.585	67.585				
2	1.500	58.068	-1	58.068			
3	1.539	55.000	-1	-1	55.000		
4	1.542	71.668	1	1	1	71.668	
5	1.543	58.060	-1	-1	1	-1	58.060
6	1.557	61.689	-1	1	1	-1	1
		Somme	**-3**	**0**	**3**	**-2**	**1**

==> | **S** | **-1**

Fig. 3.10. Tableau de calcul du τ de Kendall

Décrivons le processus de formation de S :

- Nous trions les individus selon leur taille (X). De fait, puisque nous ne gérons pas les ex aequo à ce stade, $j > i \Rightarrow x_j > x_i$.
- Pour l'individu n^o1 avec $(X = 1.496)$ et surtout $(y_1 = 67.585)$, nous regardons les individus qui sont concordants (resp.discordants) pour leur attribuer la valeur $\nu_{1j} = +1$ (resp. $\nu_{1,j}) - 1$. C'est la colonne qui vient juste après "poids (kg)" avec l'en-tête "n^o1". On observe :
  - l'individu n^o2 est discordant, en effet $y_2 = 58.068 < y_1 \rightarrow \nu_{12} = -1$
  - l'individu n^o3 est discordant, ici aussi $y_3 = 55.000 < y_1 \rightarrow \nu_{13} = -1$
  - l'individu n^o4 est concordant, en effet $y_4 = 71.668 > y_1 \rightarrow \nu_{14} = +1$
  - etc.

- Pour aboutir à la somme $\nu_1 = (-1) + (-1) + (+1) + (-1) + (-1) = -3$

– Nous faisons de même pour l'individu $n^o 2$ c.-à-d. $\nu_2 = (-1) + (+1) + (-1) + (+1) = 0$

– etc.

– Nous pouvons ainsi former la somme $S = \sum_{i=1}^{n-1} \nu_i = (-3) + 0 + (+3) + (-2) + (+1) = -1$

Le coefficient est obtenu à l'aide de l'équation 3.15

$$\hat{\tau} = \frac{2 \times (-1)}{6 \times (6-1)} = -0.0667$$

3.5.2 Test de significativité

Dès que $n > 8$ (voir [1], page 115), et plus sûrement lorsque $n > 10$ ([9], page 252), nous pouvons nous appuyer sur la normalité asymptotique de $\hat{\tau}$ sous l'hypothèse d'indépendance de X et Y.

Le test de significativité repose alors sur la statistique

$$U = \frac{\hat{\tau}}{\sqrt{\frac{2(2n+5)}{9n(n-1)}}} = 3\hat{\tau}\sqrt{\frac{n(n-1)}{2(2n+5)}} \tag{3.16}$$

U suit une loi normale centrée et réduite sous H_0. La région critique du test pour un risque α s'écrit

$$|U| > u_{1-\frac{\alpha}{2}}$$

Un exemple. L'approximation est bien évidemment mauvaise ($n = 6$) pour notre exemple ci-dessus (Figure 3.10). Nous allons quand même l'utiliser pour illustrer simplement la démarche. Rappelons que $\hat{\tau} = -0.0667$. Nous obtenons U avec

$$U = 3 \times (-0.0667) \times \sqrt{\frac{6(6-1)}{2(2 \times 6 + 5)}} = -0.1879$$

En comparant $|U|$ avec le seuil critique du test $u_{0.975} = 1.96$, nous concluons que les données sont compatibles avec l'hypothèse d'absence de lien entre X et Y.

3.5.3 Relation avec le ρ de Spearman

τ de Kendall et ρ de Spearman sont tous les deux des coefficients de corrélation de rangs. Ils reposent sur les mêmes hypothèses et exploitent les mêmes informations, il est logique qu'ils aient une puissance similaire (la capacité à détecter à juste titre l'hypothèse H_1). La différence se joue surtout sur l'interprétation des valeurs proposées par les statistiques : ρ^2 s'interprète comme une proportion de variance expliquée, à l'instar du coefficient de Pearson, τ s'interprète comme une probabilité [14].

Il y a cependant une relation entre les valeurs estimées, on montre que (voir [9], page 251) que

$$-1 \leq 3\hat{\tau} - 2\hat{\rho} \leq +1$$

14. http://www.unesco.org/webworld/idams/advguide/Chapt4_2.htm

Lorsque n est assez grand, et les coefficients pas trop proches de 1 (en valeur absolue), on constate également la relation suivante (voir [1], page 114)

$$\hat{\rho} \approx \frac{3}{2}\hat{\tau}$$

Enfin, lorsque le (X, Y) suit une loi normale bivariée, nous avons la relation (voir [7], page 138)

$$\tau = \frac{2}{\pi} \arcsin \rho$$

3.5.4 Traitement des ex-aequo

Lorsque les données comportent des ex aequo, la formule 3.15 doit être corrigée.

Calcul de ν_{ij}. Pour le calcul des écart entre paires concordantes et discordantes S, nous devons réaménager la quantité ν_{ij} en introduisant un nouveau cas : $\nu ij = 0$ si $(x_i = x_j)$ ou $(y_i = y_j)$.

Facteur de correction. Détaillons la procédure de calcul du facteur de correction E_x pour X (la démarche est identique pour E_y de Y) :

- Pour un échantillon de taille n, nous recensons les valeurs distinctes de X, elle est égale à G_x. Si $G_x = n$, il n'y a pas d'ex aequo.
- Pour chaque valeur x_g de X, nous comptabilisons le nombre d'occurrences t_g.
- Le facteur de correction E_x s'écrit alors

$$E_x = \sum_{g=1}^{G_x} t_g(t_g - 1) \tag{3.17}$$

Remarque 13 (Facteur de correction). Attention, le facteur de correction E_x est différent de celui utilisé pour le ρ de Spearman (T_x). Ici aussi, nous remarquons que $E_x = 0$ si les données ne comportent pas d'ex-aequo.

Coefficient de Kendall corrigé. Il faut maintenant introduire les facteurs de corrections pour les données comportant des ex-aequo

$$\hat{\tau} = \frac{2 \times S}{\sqrt{n(n-1) - E_x} \times \sqrt{n(n-1) - E_y}} \tag{3.18}$$

Exemple. On demande à 2 enseignants de noter de manière indépendante des dissertations de $n = 8$ étudiants. Le premier est expérimenté (X), le second est novice dans la profession (Y). On chercher à savoir si les notes attribuées sont indépendantes, auquel cas il y aurait matière à s'inquiéter concernant le degré de subjectivité que peut comporter la notation des copies.

De nouveau nous construisons le tableau de calcul sous EXCEL (Figure 3.11) :

- $n = 8$ observations.
- Nous trions les données selon les valeurs de X.

Copie	Note.Expé	Note.Novice	1	2	3	4	5	6	7
1	6.5	8.5							
2	9	8.5	0						
3	9	6.5	-1	0					
4	12	11	1	1	1				
5	12	12	1	1	1	0			
6	12	11	1	1	1	0	0		
7	13	15	1	1	1	1	1	1	
8	14	13	1	1	1	1	1	1	-1
		Somme	**4**	**5**	**5**	**2**	**2**	**2**	**-1**

==> | S | 19 |

G.x
5

Valeurs	t	Calc(E.x)
6.5	1	0
9	2	2
12	3	6
13	1	0
14	1	0
	E.x	**8**

G.y
6

Valeurs	t	Calc(E.y)
6.5	1	0
8.5	2	2
11	2	2
12	1	0
13	1	0
15	1	0
	E.y	**4**

Tau	0.76061
U	2.63483
p-value	0.00842

Fig. 3.11. Tableau de calcul du τ de Kendall en présence d'ex aequo

- Il y a $G_x = 5$ valeurs distinctes de X, nous comptons les occurrences (6.5 : 1; 9 : 2; 12 : 3; 13 : 1; 14 : 1). A l'aide de la formule 3.17, nous produisons $E_x = 8$.
- Nous procédons de la même manière pour Y. Il a $G_y = 6$ valeurs distinctes, nous obtenons $E_y = 4$.
- Il faut maintenant produire la valeur de S. Nous prenons comme référence l'individu n^o1 avec $(x_1 = 6.5; y_1 = 8.5)$. Regardons les paires concordantes et discordantes :
  - n^o2 est ex-aequo, en effet $y_2 = y_1 \rightarrow \nu_{12} = 0$
  - n^o3 est discordant car $y_2 = 6.5 < y_1 \rightarrow \nu_{13} = -1$
  - Etc.
- Pour l'individu n^o1, nous obtenons ainsi $\nu_1 = 4$
- Prenons maintenant comme référence l'individu n^o2 avec $(x_2 = 9; y_2 = 8.5)$
  - n^o3 est ex-aequo car $x_2 = 9 = x_1 \rightarrow \nu_{23} = 0$; il n'est même pas nécessaire de considérer la valeur de Y pour cette paire.
  - n^o4 est concordant car $x_4 = 12 > x_2$ et $y_4 = 11 > y_2 \rightarrow \nu_{24} = +1$
  - Etc.
- Pour l'individu n^o2, nous obtenons $\nu_2 = 5$
- Etc. Pour aboutir au final à $S = 19$

Nous utilisons la formule corrigée (Equation 3.18)

$$\hat{\tau} = \frac{2 \times 19}{\sqrt{8(8-1)-8} \times \sqrt{8(8-1)-4}} = 0.76061$$

Pour tester la significativité du coefficient, nous utilisons l'approximation normale

$$U = 3 \times 0.76061 \sqrt{\frac{8(8-1)}{2(2 \times 8 + 5}} = 2.63483$$

La p-value est 0.00842. Au risque 5%, on peut conclure à l'existence d'un lien positif entre un correcteur expérimenté et un correcteur novice. Mieux même, puisque nous pouvons interpréter le τ de

Kendall comme une probabilité, nous dirions que 76.06% correspond au surcroît de chances que les deux correcteurs rangent de la même manière 2 copies prises au hasard (ouf!).

3.6 Rapport de corrélation

Lorsque la relation s'écarte de la linéarité, nous constatons que le coefficient de corrélation n'est plus adapté, particulièrement lorsque la relation est non monotone. Dans cette section, nous présentons un indicateur, le rapport de corrélation [15], dont l'interprétation et l'efficacité ne dépend pas de la forme de la relation étudiée. En particulier, il permet de rendre compte de la liaison même si elle est non monotone.

3.6.1 Principe et interprétation

Le rapport de corrélation [16] est une **mesure asymétrique**, elle repose sur la notion d'**espérance conditionnelle**. Nous notons $E[Y/X = x]$ l'espérance de la variable Y lorsque $X = x$, elle nous fournit un résumé de Y lorsque X prend la valeur x. Dans la régression linéaire simple par exemple, nous faisons l'hypothèse que cette espérance est une fonction linéaire de X c.-à-d. $E[Y/X = x] = a \times X + b$.

Dans le cas du rapport de corrélation, nous estimons directement cette quantité à partir des observations. Cela suppose, et c'est la principale limite de cette mesure, que l'on dispose de plusieurs observations de Y pour chaque valeur x de X.

Le **rapport de corrélation théorique** $\eta^2_{Y/X}$ est définie comme le rapport entre la variabilité de Y expliquée par X et la variabilité totale de Y.

$$\eta^2_{y/x} = \frac{E\{(Y/X - E[Y])^2\}}{E\{(Y - E[Y])^2\}} \tag{3.19}$$

Domaine de définition. Le rapport de corrélation [17] $\eta^2_{y/x}$ est défini sur l'intervalle $0 \leq \eta^2_{y/x} \leq 1$.

- Lorsqu'il est égal à 0, cela veut dire que la connaissance de X ne donne aucune information sur Y. La moyenne de Y est la même quelle que soit la valeur de X.
- A contrario, lorsqu'il est égal à 1, la connaissance de X permet de déterminer avec certitude la valeur de Y c.-à-d. à chaque valeur x de X correspond une seule valeur de Y.

Liaison entre une variable qualitative et une variable quantitative. Le rapport de corrélation a une portée plus large que la simple alternative pour mesurer une liaison non linéaire entre 2 variables quantitatives. Nous constatons dans la définition ci-dessus (formule 3.19) qu'à aucun moment nous faisons référence au caractère ordonné de X. De fait, le rapport de corrélation peut être utilisé pour caractériser l'association entre une variable qualitative X et une variable quantitative Y ([7], page 143). On se rapproche en cela du schéma de l'analyse de variance (ANOVA).

15. en anglais, *coefficient of nonlinear relationship*, ou *eta coefficient*, ou encore *eta correlation ratio*
16. Voir `http://biblioxtrn.uqar.qc.ca/stat/Fichesstat/multivariable/quanti/rapport.htm`
17. Voir `http://nte-serveur.univ-lyon1.fr/nte/immediato/math2002/Mass11/cours/chapitr3d.htm`

Si X prend K valeurs distinctes, calculé sur un échantillon, le **rapport de corrélation empirique** est définie de la manière suivante :

$$\hat{\eta}^2_{y/x} = \frac{\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.20}$$

avec n_k le nombre d'observations telles que $X = x_k$, $\bar{y}_k$ la moyenne de Y lorsque $X = x_k$.

Nous pouvons aussi écrire le rapport de corrélation en faisant intervenir la variance de Y non expliquée par les X c.-à-d. la variance résiduelle.

$$\hat{\eta}^2_{y/x} = 1 - \frac{\sum_{k=1}^{K} n_k \sum_{i=1}^{n_k}(y_i - \bar{y}_k)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.21}$$

La formule n'utilise jamais de manière explicite les valeurs x_k. De même, elle ne tient pas compte du caractère ordonné de X c.-à-d. $x_{k+1} > x_k$. On fait donc l'impasse sur une information qui est pourtant importante. C'est le prix à payer pour ne pas avoir à faire d'hypothèses sur la forme de la relation.

On voit bien la limite de l'indicateur dans cette nouvelle formulation. Si nous ne disposons que d'une seule observation pour chaque valeur de X c.-à-d. $K = n$, $n_k = 1$, $\forall k$ et $y_i = \bar{y}_k$. Le rapport de corrélation sera mécaniquement égal à 1 sans qu'il n'y ait aucune relation entre X et Y. Néanmoins cette restriction n'est pas aussi contraignante qu'on pourrait le penser :

1. Dans les sciences expérimentales où les données sont le fruit d'une expérimentation raisonnée, la répétition des observations pour une valeur de X est tout à fait naturelle. Par exemple, pour évaluer la réduction du nombre de microbes consécutive à l'administration d'un médicament, on répartit les cobayes en groupes, dans un groupe on donne une dose identique. Nous disposons de plusieurs valeurs de Y (réduction des microbes) pour chaque valeur de X (dose du médicament).
2. Nous avons la possibilité de découper les valeurs de X en classes de manière à obtenir un certain nombre d'observations dans chaque groupe. Dans ce cas, le choix des bornes des intervalles est déterminant. Si elles sont mal définies, des informations primordiales peuvent être masquées. A l'extrême, si on ne prend qu'un seul intervalle qui va du minimum au maximum, on ne pourra rien en tirer.

Relations entre coefficient de corrélation et rapport de corrélation. Ces deux indicateurs sont censés mesurer le lien entre deux variables, à la différence que le premier fait l'hypothèse de la linéarité de la relation. On peut noter alors quelques relations importantes entre r^2_{xy} et $\eta^2_{y/x}$:

- De manière générale, $\eta^2_{y/x} \geq r^2_{xy}$. On le comprend aisément, r introduit une contrainte supplémentaire, l'hypothèse de linéarité, pour mesurer la liaison. On peut d'ailleurs utiliser l'écart $(\eta^2_{y/x} - r^2_{xy})$ pour évaluer le caractère linéaire de la relation.
- $r^2 = 1 \Rightarrow \eta^2 = 1$, une liaison linéaire parfaite signifie une liaison parfaite.
- $\eta^2 = 0 \Rightarrow r^2 = 0$, absence totale de liaison implique absence de liaison linéaire.
- La valeur x_k n'entre pas en ligne de compte dans le calcul du rapport de corrélation. Si on la remplace artificiellement par la moyenne conditionnelle de Y c.-à-d. $x_k = \bar{y}_k$, alors $\hat{\eta}^2_{y/x} = \hat{r}^2_{xy}$ (voir [7], page 144).

3.6.2 Inférence statistique

Pour tester la significativité du rapport de corrélation, il faut se référer au schéma d'analyse de variance à 1 facteur [18]. En effet, le test d'hypothèses [19]

$$H_0 : \eta^2_{y/x} = 0$$
$$H_1 : \eta^2_{y/x} > 0$$

Est équivalent à [20]

$$H_0 : \mu_1 = \cdots = \mu_K$$
$$H_1 : \text{une au moins differe des autres}$$

Sous l'hypothèse nulle, et sous condition que les distributions conditionnelles soient gaussiennes et de variance identique (hypothèse d'*homoscédasticité*) [21], la statistique :

$$F = \frac{\frac{\hat{\eta}^2}{K-1}}{\frac{1-\hat{\eta}^2}{n-K}} = \frac{n-K}{K-1} \times \frac{\hat{\eta}^2}{1-\hat{\eta}^2} \tag{3.22}$$

Suit une loi de Fisher à $(K-1, n-K)$ degrés de liberté.

Pour un risque α, la région critique du test s'écrit :

$$R.C. : F > F_{1-\alpha}(K-1, n-K)$$

où $F_{1-\alpha}(K-1, n-K)$ est le quantile d'ordre $(1-\alpha)$ de la loi de Fisher à $(K-1, n-K)$ degrés de liberté.

3.6.3 Un exemple

Nous essayons de vérifier, au risque de 10%, l'influence de la consommation de cigarettes (en nombre de paquets par jour) sur le risque d'apparition de la leucémie chez 43 gros fumeurs. L'analyse est bien asymétrique, dans l'autre sens, *a priori*, elle n'aurait pas trop d'intérêt [22].

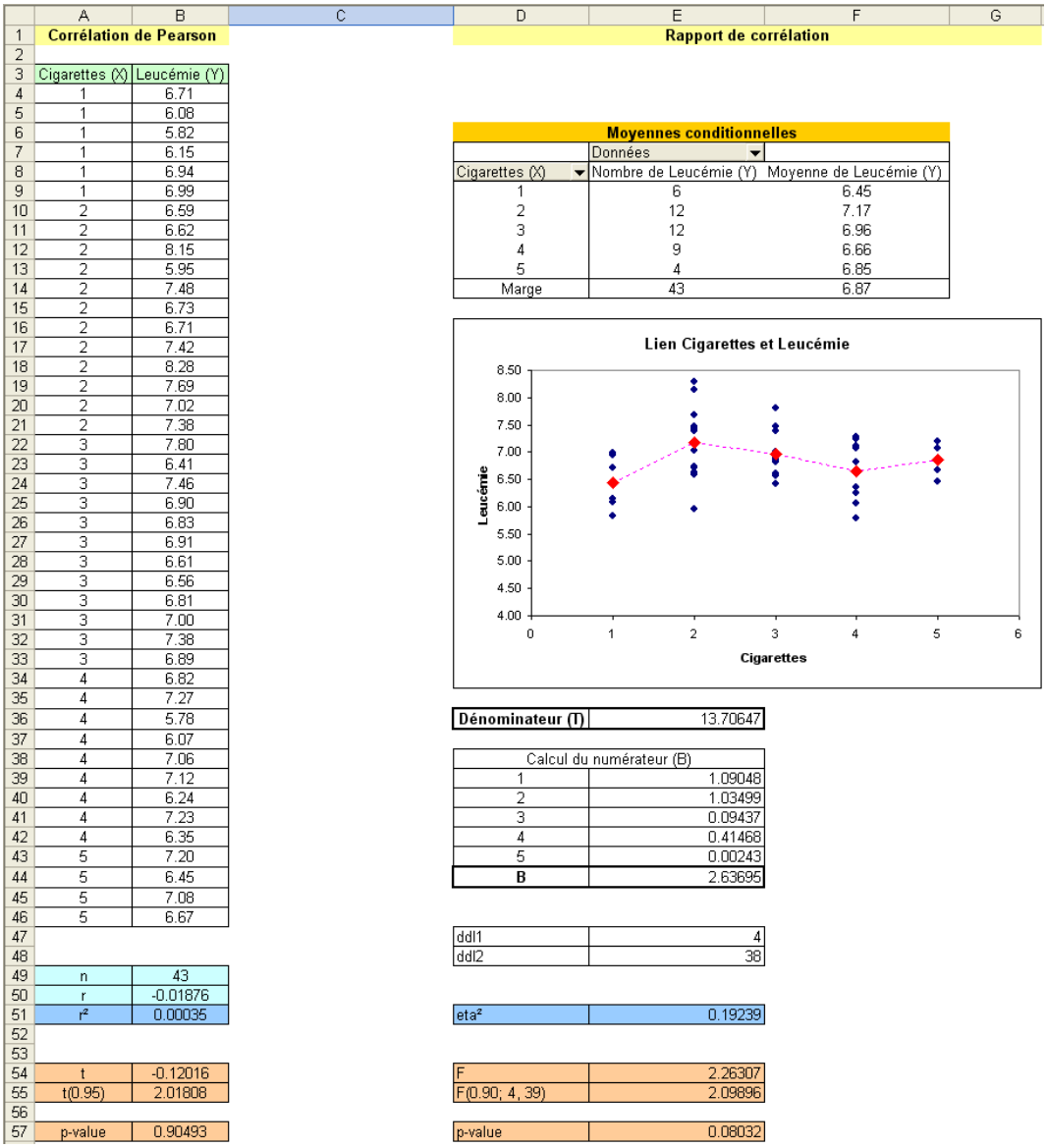
A partir de ces $n = 43$ observations, nous menons dans un premier temps une analyse de corrélation classique en calculant le coefficient de Pearson (Figure 3.12, colonnes A et B de la feuille de calcul). Nous obtenons :

- Le coefficient de corrélation empirique est $\hat{r} = -0.01876$, son carré $\hat{r}^2 = 0.00035$
- Pour tester la significativité, nous formons le t de Student, $t = -0.12016$

18. `http://spiral.univ-lyon1.fr/mathsv/cours/pdf/stat/Chapitre9.pdf`
19. Le rapport de corrélation est toujours positif ou nul, le test est forcément unilatéral.
20. $\mu_k = E[Y/X = x_k]$, la moyenne conditionnelle théorique
21. l'ANOVA est quand même bien robuste par rapport à ces hypothèses
22. Les données sont fictives, que le lecteur médecin ne s'affole pas.

– La p-value du test nous fournit p-value = 0.90493

– Au risque de 10%, il semble patent qu'il n'y a aucun lien entre les deux variables. On peut fumer en paix.



Corrélation de Pearson	
Cigarettes (X)	Leucémie (Y)
1	6.71
1	6.08
1	5.82
1	6.15
1	6.94
1	6.99
2	6.59
2	6.62
2	8.15
2	5.95
2	7.48
2	6.73
2	6.71
2	7.42
2	8.28
2	7.69
2	7.02
2	7.38
3	7.80
3	6.41
3	7.46
3	6.90
3	6.83
3	6.91
3	6.61
3	6.56
3	6.81
3	7.00
3	7.38
3	6.89
4	6.82
4	7.27
4	5.78
4	6.07
4	7.06
4	7.12
4	6.24
4	7.23
4	6.35
5	7.20
5	6.45
5	7.08
5	6.67
n	43
r	-0.01876
r^2	0.00035
t	-0.12016
t(0.95)	2.01808
p-value	0.90493

Rapport de corrélation		
Moyennes conditionnelles		
	Données	
Cigarettes (X)	Nombre de Leucémie (Y)	Moyenne de Leucémie (Y)
1	6	6.45
2	12	7.17
3	12	6.96
4	9	6.66
5	4	6.85
Marge	43	6.87

Dénominateur (I)	13.70647

Calcul du numérateur (B)	
1	1.09048
2	1.03499
3	0.09437
4	0.41468
5	0.00243
B	2.63695

ddl1	4
ddl2	38

eta^2	0.19239

F	2.26307
F(0.90; 4, 39)	2.09896
p-value	0.08032

Fig. 3.12. Rapport de corrélation - Risque de leucémie vs. Consommation de cigarettes

S'arrêter à ce stade serait une grave erreur, un petit graphique mettant en relation les deux variables éclaire la relation sous un autre jour. Calculons maintenant le rapport de corrélation (Figure 3.12, colonnes D à F de la feuille de calcul) :

– Dans le graphique, on se rend compte que pour chaque valeur de X, les nuages de points correspondant sont assez décalés. Impression confirmée par les moyennes conditionnelles en rouge que nous

avons reliées. S'il y avait eu absence de relation, les moyennes seraient au même niveau, nous aurions obtenu un droite horizontale. Il semble que ce ne soit pas le cas ici, vérifions cela numériquement.

- Pour calculer le rapport de corrélation, nous devons tout d'abord former les moyennes conditionnelles, nous avons réalisé cela à l'aide de l'outil "tableaux croisés dynamiques" d'EXCEL, nous avons à la fois les effectifs et les moyennes par valeur de X. Par exemple, pour $X = 1$, nous avons $n_1 = 6$ et $\bar{y}_1 = 6.45$
- L'effectif global est bien $n = 43$ et la moyenne $\bar{y} = 6.87$.
- Nous calculons le numérateur de la formule 3.20, nous obtenons $B = 2.63695$
- De la même manière, nous formons le dénominateur, nous obtenons $T = 13.70647$
- Le rapport de corrélation estimé est égal à $\hat{\eta}^2 = \frac{B}{T} = 0.19239$. A comparer avec $\hat{r}^2 = 0.00035$ obtenu précédemment. Si liaison il y a, elle n'est absolument pas linéaire en tous les cas.
- Voyons justement ce qu'il en est de la significativité. Nous formons la statistique F (équation 3.22), elle est égale à $F = 2.26307$.
- Pour un risque $\alpha = 0.1$, nous la comparons à $F_{0.9}(4, 38) = 2.09896$. Au risque $\alpha = 10\%$, le rapport de corrélation est différent de 0, résultat confirmé par la p-value égale à 0.08032.
- Il y a donc bien un lien entre la consommation de cigarettes et le risque de leucémie, mais la liaison est assez complexe. On a des sérieux problèmes quand on en consomme 2 paquets par jour, au delà, on dirait que la situation s'améliore (*ah bon ? !*). Mais il ne faut pas se faire d'illusions, à mon avis, c'est parce qu'on va mourir d'autre chose *avant* de contracter une leucémie.

Partie II

Corrélations partielles et semi-partielles

4

Corrélation partielle paramétrique et non paramétrique

4.1 Principe de la corrélation partielle

Il n'est pas rare qu'une ou plusieurs autres variables viennent fausser la corrélation entre 2 variables, laissant à penser à tort l'existence (ou l'absence) d'une liaison. On parle de facteur confondant (voir section 2.6, *Problèmes et cas pathologiques*). La littérature statistique regorge d'exemples plus ou moins loufoques de corrélations numériquement élevées, mais qui ne résistent pas une seconde à l'interprétation :

- Corrélation entre les ventes de lunettes noires et les ventes de glaces (c'est pour ne pas voir les calories qu'on engouffre...). Il faut surtout y voir l'effet de la chaleur ou de l'ensoleillement.
- Corrélation entre le nombre d'admissions à l'hôpital et les ventes de glaces (ça y est, les calories ont encore frappé...). Encore une fois, la canicule y est pour quelque chose peut être.
- Corrélation entre la longueur des cheveux et la taille des personnes (et oui, on compense comme on peut...). On a mélangé les hommes et les femmes dans les données. En moyenne, les hommes sont plus grands que les femmes avec, *a contrario*, des cheveux plus courts (Figure 2.9).
- Corrélation entre le prix des voitures et leur consommation (tant qu'à payer, autant le faire *ad vitam* ...). Les voitures luxueuses, chères, sont aussi souvent de lourdes grosses cylindrées. Toute la filière automobile vous dit merci.
- Corrélation entre la hausse des prix et le budget alimentation des ménages (les soucis donnent faim, c'est bien connu...). Il faudrait plutôt exprimer la consommation alimentaire en volume, autrement en tous les cas.
- Etc.

L'idée de la corrélation partielle est de mesurer la corrélation entre X et Y en annulant (en contrôlant) l'effet d'une troisième variable Z. Lorsque cette dernière est qualitative, la stratégie est simple, il s'agit de calculer $\hat{r}$ dans chaque groupe du point de vue numérique, et de distinguer explicitement les groupes dans le graphique nuage de points (Figure 2.9 par exemple pour la corrélation taille et longueur de cheveux).

L'affaire se complique lorsque la **variable de contrôle** Z est elle aussi numérique [1]. Il faudrait alors retrancher de X et Y la variance expliquée par Z, puis calculer la corrélation en utilisant l'information résiduelle. C'est exactement la démarche de la corrélation partielle.

Le rôle de Z est complexe. Parfois elle exacerbe la corrélation entre X et Y, parfois elle la masque. Garson ([5], `http://www2.chass.ncsu.edu/garson/pa765/partialr.htm`) résume dans un graphique les différentes interaction qu'il peut y avoir entre X, Y et Z (Figure 4.1).

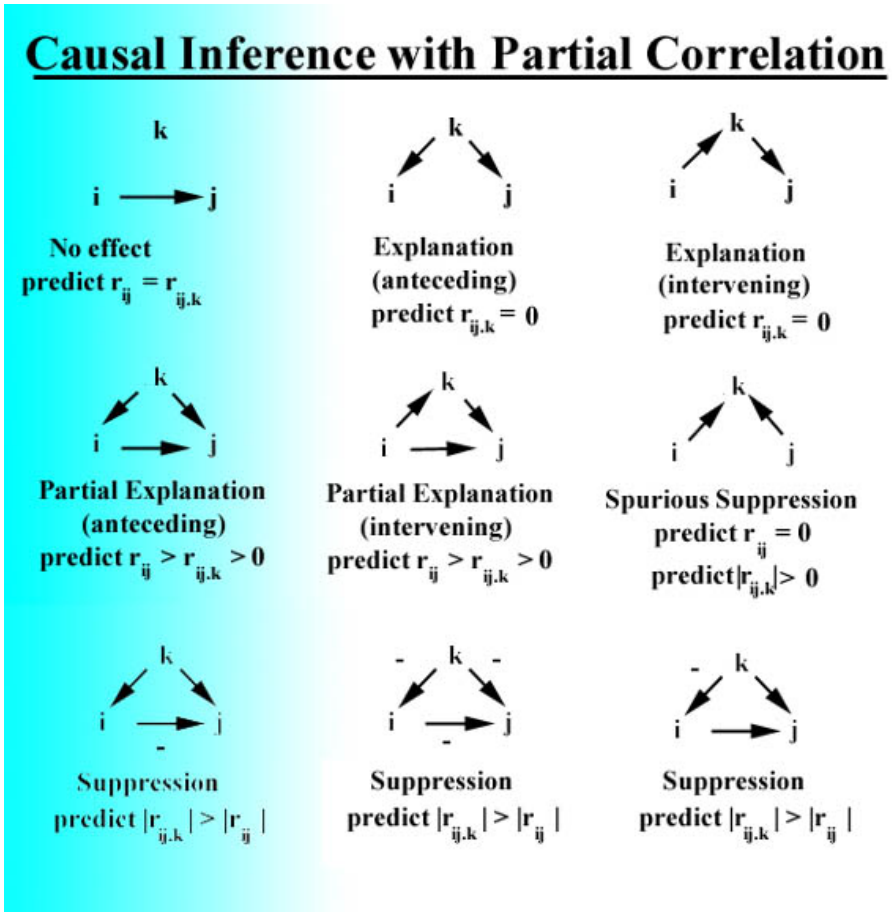


Fig. 4.1. Typologie de l'influence de Z sur la corrélation r_{xy}

On parle de corrélation brute lorsque l'on souhaite mesurer la relation directe r_{xy}. On parle de corrélation partielle lorsque l'on souhaite faire intervenir une ou plusieurs variables de contrôle : plus précisément, corrélation partielle d'ordre p lorsque l'on a p variables de contrôle.

1. Dans les sciences expérimentales où nous contrôlons la production des données, nous pourrions, pour chaque valeur de Z, répéter l'expérimentation de manière à recueillir plusieurs observations (x_i, y_i). On retrouve ainsi le schéma de la variable de contrôle discrète. Mais dans les sciences sociales, souvent le triplet (x_i, yi, z_i) est unique dans le fichier, la seule solution est de passer par la corrélation partielle.

Corrélation (même partielle) n'est toujours pas causalité. Précisons encore et toujours qu'il s'agit toujours là de procédures numériques destinées à mesurer l'existence et l'intensité d'une liaison. La corrélation partielle ne déroge pas à cette règle. La mise en évidence d'une éventuelle causalité ne peut et ne doit reposer que sur les connaissances du domaine. En revanche, et c'est pour cela qu'elle peut être très bénéfique dans une analyse, la corrélation partielle peut permettre de clarifier la relation qui existe (ou qui n'existe pas) entre 2 variables.

Remarque 14 (Quelques éléments sur les notations). Dans cette partie du support, nous noterons en priorité r le coefficient partiel, sauf s'il y a ambiguïté, auquel cas nous indiquerons les indices adéquats. Concernant la transformation de Fisher, pour éviter la confusion avec la (ou les) variable(s) de contrôle, nous la noterons f.

4.2 Corrélation partielle d'ordre 1 basé sur le r de Pearson

Dans un premier temps, étudions le coefficient de corrélation partielle d'ordre 1 basée sur le coefficient de Pearson. Les hypothèses relatives à l'inférence statistique restent de mise ici, on postule notamment que la distribution de (X, Y) conditionnellement à Z suit une loi normale bivariée (voir [7], page 133). Fort heureusement, les propriétés asymptotiques sont conservées. Il n'en reste pas moins que le coefficient partiel ne caractérise que les relations linéaires.

4.2.1 Définition - Estimation

La consommation partielle $r_{xy.z}$ peut être définie à partir des corrélations brutes

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{1 - r_{xz}^2} \times \sqrt{1 - r_{yz}^2}} \tag{4.1}$$

L'idée est assez limpide, on retranche de la relation directe (X, Y) les relations respectives de X et Y avec Z. Puis un terme de normalisation (symétrique, X vs. Z et Y vs. Z) est introduit de manière à ce que $-1 \leq r_{xy.z} \leq +1$

Remarquons plusieurs résultats intéressants. Pour fixer les idées, sans que cela ne réduise la portée du propos, nous dirons que $r_{xy} > 0$:

- Lorsque Z est indépendant de X et Y ($r_{xz} = r_{yz} = 0$), $r_{xy.z} = r_{xy}$ c.-à-d. Z ne pèse en aucune manière dans la relation entre X et Y
- Lorsque Z est fortement lié positivement avec X et Y, on peut aboutir au résultat $r_{xy.z} \approx 0$ c.-à-d. il n'y a rien dans la relation (X, Y) qui ne soit pas déjà expliquée par Z
- Lorsque les liaisons entre Z d'une part, X et Y d'autre part, sont de signe opposés (ex. $r_{xz} > 0$ et $r_{yz} < 0$), le produit $r_{xz}.r_{yz} < 0$, on constate que $r_{xy.z} > r_{xy}$

L'**estimation** de la corrélation partielle passe simplement par l'introduction des estimations des corrélations brutes dans la formule 4.1 c.-à-d.

$$\hat{r}_{xy.z} = \frac{\hat{r}_{xy} - \hat{r}_{xz}\hat{r}_{yz}}{\sqrt{1-\hat{r}_{xz}^2} \times \sqrt{1-\hat{r}_{yz}^2}} \tag{4.2}$$

4.2.2 Exemple des voitures

Reprenons notre exemple des voitures (Figure 2.2). Nous souhaitons clarifier la liaison entre "puissance" (X) et "consommation" (Y) en contrôlant le rôle de la cylindrée (Z). En effet, une grosse cylindrée a tendance à être puissante, mais elle a tendance aussi à consommer plus que de raison : au final, que reste-t-il de la liaison (Y, X) une fois que l'on a retranché l'explication (en termes de variance) fournie par Z ?

	C	D	E
1	X	Y	Z
2	Puissance	Conso	Cylindree
3	32	5.7	846
4	39	5.8	993
5	29	6.1	899
6	44	6.5	1390
7	33	6.8	1195
8	32	6.8	658
9	55	7.1	1331
10	74	7.4	1597
11	74	9.0	1761
12	101	11.7	2165
13	85	9.5	1983
14	85	9.5	1984
15	89	8.8	1998
16	65	9.3	1580
17	54	8.6	1390
18	66	7.7	1396
19	106	10.8	2435
20	55	6.6	1242
21	107	11.7	2972
22	150	11.9	2958
23	122	10.8	2497
24	66	7.6	1998
25	125	11.3	2496
26	89	10.8	1998
27	92	9.2	1997
28	85	11.6	1984
29	97	12.8	2438
30	125	12.7	2473

	H	I
	n	28

Corrélations brutes		
Puissance	Conso	0.88781
Puissance	Cylindrée	0.94755
Conso	Cylindrée	0.89187

Corrélation partielle	
r_xy.z	0.29553

Test de significativité	
t	1.54673
t(0.975 ; 25)	2.38461
p-value	0.13450

Intervalle de confiance à 95%	
f	0.30461
e.t.	0.20000
u(0.975)	1.95996
bb(f)	-0.08738
bh(f)	0.69661
bb (r)	-0.08716
bh (r)	0.60221

Fig. 4.2. Calcul de la corrélation partielle d'ordre 1 - Fichier "voitures"

Détaillons les calculs de la feuille EXCEL (Figure 4.2) :

- Nous calculons les corrélations brutes $r_{xy} = 0.88781$, $r_{xz} = 0.94755$ et $r_{yz} = 0.89187$. D'ores et déjà, nous constatons que la variable de contrôle est fortement liée avec X et Y.
- Appliquons la formule 4.2 sur ces corrélations, nous obtenons

$$r_{xy.z} = \frac{0.88781 - 0.94755 \times 0.89187}{\sqrt{(1 - 0.94755^2)(1 - 0.89187^2)}} = 0.29553$$

– La corrélation partielle est singulièrement réduite si l'on se réfère à la corrélation brute. Apparemment, "cylindrée" joue beaucoup dans la liaison entre "puissance" et "consommation". Nous essaierons de voir dans la section suivante si, néanmoins, la relation résiduelle reste significative.

4.2.3 Test de significativité et intervalle de confiance

Test de significativité. Si l'hypothèse de normalité est vérifiée, le test de significativité équivaut à un test d'indépendance c.-à-d. X est indépendant de Y conditionnellement à Z. Dans le cas contraire, avec les propriétés asymptotiques, le test permet quand même d'éprouver la nullité du coefficient.

L'hypothèse nulle du test, qui peut être bilatéral ou unilatéral, s'écrit

$$H_0 : r_{xy.z} = 0$$

Sous H_0, la statistique du test

$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-3}}} \tag{4.3}$$

suit une loi de Student à $(n-3)$ degrés de liberté.

Pour un risque α et un test bilatéral, nous rejetons l'hypothèse nulle si

$$R.C. : |t| > t_{1-\alpha/2}(n-3)$$

où $t_{1-\alpha/2}(n-3)$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $(n-3)$ degrés de liberté.

Revenons à notre exemple numérique (Figure 4.2), Nous calculons :

– les degrés de liberté $n - 3 = 28 - 3 = 25$;
– la statistique $t = 1.54673$;
– dont la valeur absolue est comparée avec le seuil critique $t_{0.975}(25) = 2.38461$
– Au risque $\alpha = 5\%$, nous concluons que la corrélation entre "consommation" et "puissance" conditionnellement à "cylindrée" n'est pas significativement différente de 0. En d'autres termes, à cylindrée égale, la consommation ne varie pas avec la puissance.

Intervalle de confiance. La distribution du test est uniquement valide dans le voisinage $r_{xy.z} = 0$. Pour élaborer l'intervalle de confiance au niveau $(1-\alpha)$, nous devons passer, comme pour la corrélation brute, par la transformation de Fisher.

Elle est définie de la même manière,

$$f = \frac{1}{2} \ln \frac{1 + r_{xy.z}}{1 - r_{xy.z}}$$

L'estimateur $\hat{f}$ est calculée à l'aide de l'estimation de la corrélation partielle, il est asymptotiquement sans biais, distribué selon une loi normale, et de variance [2]

$$\sigma_f^2 = \frac{1}{n-3} \tag{4.4}$$

Dans notre exemple (Figure 4.2), nous souhaitons construire l'intervalle de confiance à 95% :

- Nous calculons la transformation de Fisher $f = \frac{1}{2} \ln \frac{1+0.29553}{1-0.29553} = 0.30461$
- L'écart type associé est égale à $\sigma_f = \sqrt{\frac{1}{28-3}} = 0.2$
- Le quantile d'ordre 975% est $u_{0.975} = 1.95996$
- La borne basse (resp. haute) pour f est $bb_f = 0.30461 - 1.95996 \times 0.2 = -0.08738$ (resp. $bh_f = 0.30461 + 1.95996 \times 0.2 = 0.69661$)
- Il ne reste plus qu'à appliquer la transformation inverse pour obtenir la borne basse (resp. haute) du coefficient $bb_r = \frac{e^{2\times(-0.08738)}-1}{e^{2\times(-0.08738)}+1} = -0.08716$ (resp. $bh_r = 0.60221$).
- Nous constatons que l'intervalle englobe la valeur 0, c'est une autre manière de détecter la non-significativité de r.

4.3 Corrélation partielle d'ordre p ($p > 1$) basé sur le r de Pearson

4.3.1 Définition

La corrélation partielle d'ordre p est une généralisation de la corrélation partielle. L'objectif est d'introduire plusieurs variables de contrôle. Dans notre exemple des voitures (Figure 2.2), nous savons pertinemment que le "poids" est un aspect important que la consommation. Nous souhaitons également annuler son éventuelle action dans la relation "consommation" - "puissance".

Comment estimer la corrélation partielle $r_{xy.z_1z_2...z_p}$?

Calcul récursif

On montre qu'il est possible de calculer les corrélations partielles d'ordre $p+1$ à partir des corrélations partielles d'ordre p. On utilise pour cela la formule de passage suivante, qui n'est pas sans rappeler d'ailleurs le passage des corrélations brutes vers la corrélation partielle d'ordre 1

$$r_{xy.z_1...z_pz_{p+1}} = \frac{r_{xy.z_1z_2...z_p} - r_{xz_{p+1}.z_1z_2...z_p} \times r_{yz_{p+1}.z_1z_2...z_p}}{\sqrt{1 - r^2_{xz_{p+1}.z_1z_2...z_p}} \times \sqrt{1 - r^2_{yz_{p+1}.z_1z_2...z_p}}} \tag{4.5}$$

Pour la corrélation partielle d'ordre 2 que nous mettrons en oeuvre sur un exemple ci-dessous, la formulation adéquate est

2. voir http://en.wikipedia.org/wiki/Partial_correlation; http://www.stat.psu.edu/online/development/stat505/07_partcor/06_partcor_partial.html

$$r_{xy.z_1z_2} = \frac{r_{xy.z_1} - r_{xz_2.z_1} \times r_{yz_2.z_1}}{\sqrt{1 - r^2_{xz_2.z_1}} \times \sqrt{1 - r^2_{yz_2.z_1}}} \tag{4.6}$$

Si l'écriture est simple, le calcul est assez complexe. En effet, pour obtenir la corrélation partielle d'ordre p, nous devons dans un premier temps calculer les corrélations brutes de toutes les variables 2 à 2 à partir des données c.-à-d. $\binom{p+1}{2}$ corrélations. Puis mettre à jour de proche en proche cette matrice de corrélation en introduisant la première variable de contrôle z_1, puis la seconde z_2, etc. jusqu'à ce qu'on obtienne la profondeur souhaitée.

Exemple : Mesurer la relation "puissance (X) - consommation (Y)" en contrôlant "cylindrée" (Z_1) et "poids" (Z_2) - Approche n^o1. Corsons notre affaire de voitures en introduisant 2 variables de contrôle. Nous voulons produire le résultat à partir de l'équation 4.6. La séquence des calculs est la suivante (Figure 4.3) :

		X	Y	Z1	Z2
Numero	Modele	Conso	Puissance	Cylindree	Poids
1	Daihatsu Cuore	5.7	32	846	650
2	Suzuki Swift 1.0 GLS	5.8	39	993	790
3	Fiat Panda Mambo L	6.1	29	899	730
4	VW Polo 1.4 60	6.5	44	1390	955
5	Opel Corsa 1.2i Eco	6.8	33	1195	895
6	Subaru Vivio 4WD	6.8	32	658	740
7	Toyota Corolla	7.1	55	1331	1010
8	Opel Astra 1.6i 16V	7.4	74	1597	1080
9	Peugeot 306 XS 108	9.0	74	1761	1100
10	Renault Safrane 2.2. V	11.7	101	2165	1500
11	Seat Ibiza 2.0 GTI	9.5	85	1983	1075
12	VW Golt 2.0 GTI	9.5	85	1984	1155
13	Citroen ZX Volcane	8.8	89	1998	1140
14	Fiat Tempra 1.6 Liberty	9.3	65	1580	1080
15	Fort Escort 1.4i PT	8.6	54	1390	1110
16	Honda Civic Joker 1.4	7.7	66	1396	1140
17	Volvo 850 2.5	10.8	106	2435	1370
18	Ford Fiesta 1.2 Zetec	6.6	55	1242	940
19	Hyundai Sonata 3000	11.7	107	2972	1400
20	Lancia K 3.0 LS	11.9	150	2958	1550
21	Mazda Hachtback V	10.8	122	2497	1330
22	Mitsubishi Galant	7.6	66	1998	1300
23	Opel Omega 2.5i V6	11.3	125	2496	1670
24	Peugeot 806 2.0	10.8	89	1998	1560
25	Nissan Primera 2.0	9.2	92	1997	1240
26	Seat Alhambra 2.0	11.6	85	1984	1635
27	Toyota Previa salon	12.8	97	2438	1800
28	Volvo 960 Kombi aut	12.7	125	2473	1570

n	28			
r	0.25309			
Test de siginificativité				
t	1.28161			
t(0.975; 24)	2.39095			
p-value	0.21222			
Intervalle de confiance				
f	0.25871			
e.t	0.20412			
u(0.975)	1.95996			
bb(f)	-0.14136			
bh(f)	0.65879			
bb (r)	-0.14043			
bh (r)	0.57756			

Corrélations brutes croisées

	Conso	Puissance	Cylindree	Poids
Conso	1	0.8878	0.8919	0.9263
Puissance		1	0.9475	0.8433
Cylindree			1	0.8616
Poids				1

Corrélations partielles / Z1 (cylindrée)

	Conso	Puissance	Cylindree	Poids
Conso	1	0.2955	-	0.6878
Puissance		1	-	0.1663
Cylindree			1	-
Poids				1

Corrélations partielles / Z1,Z2 (cylindrée, poids)

	Conso	Puissance	Cylindree	Poids
Conso	1	**0.25309**	-	-
Puissance		1	-	-
Cylindree			1	-
Poids				1

Fig. 4.3. Corrélation partielle d'ordre 2 - Approche récursive - Fichier "voitures"

- Tout d'abord nous calculons les corrélations brutes croisées : $r_{xy} = 0.8878$, $r_{xz_1} = 0.8819$, $r_{xz_2} = 0.9263$, etc. C'est l'objectif de la matrice "Corrélations brutes croisées" dans la partie basse de la feuille EXCEL.
- Ensuite, nous devons calculer toutes les corrélations croisées d'ordre 1 où Z_1 (cylindrée) joue le rôle de variable de contrôle. Nous obtenons $r_{xy.z_1} = 0.2955$, $r_{xz_2.z_1} = 0.6878$ et $r_{yz_2.z_1} = 0.1663$ (cf. la matrice "Corrélations partielles / Z1")
- Enfin, dernière étape, à partir de la matrice précédente nous appliquons l'équation 4.6 pour introduire la seconde variable de contrôle Z_2 (poids). Nous obtenons

$$r_{xy.z_1z_2} = \frac{0.2955 - 0.6878 \times 0.1663}{\sqrt{1 - 0.6878^2} \times \sqrt{1 - 0.1663^2}} = 0.25309$$

  Il n'y a plus qu'un seul chiffre dans la matrice "Corrélations partielles /Z1,Z2", nous sommes arrivés au bout du processus récursif.

Tant que le nombre de variables reste faible, ce processus est intéressant, surtout pédagogiquement. Lorsqu'il devient élevé, nous utilisons une autre approche, plus efficace, plus directe, pour obtenir la valeur de la corrélation partielle d'ordre p.

Calcul par les résidus de la régression

Cette approche s'appuie sur un autre point de vue pour aboutir au même résultat. Rappelons que la corrélation partielle consiste à mesurer le lien entre l'information résiduelle de X et Y qui ne soit pas déjà expliquée par les variables de contrôle. En prenant au pied de la lettre cette description, on s'attache à calculer le résidu e_x (resp. e_y) de la régression de X (resp. Y) sur $(Z_1, Z_2, \ldots, Z_p)$. Estimer la corrélation partielle d'ordre p revient tout simplement à calculer la corrélation brute entre les résidus

$$\hat{r}_{xy.z_1 \ldots z_p} = \hat{r}_{e_x e_y} \tag{4.7}$$

Exemple : Mesurer la relation "puissance (X) - consommation (Y)" en contrôlant "cylindrée" (Z_1) et "poids" (Z_2) - Approche n^o2

La feuille de calcul est organisée de manière différente maintenant (Figure 4.4).

- Tout d'abord, nous devons produire les équations de régression, nous obtenons $\hat{X} = 0.00443Z_2 + 0.00130Z_1 + 1.14755$. Nous en déduisons la nouvelle colonne de résidus $e_x = X - \hat{X}$ (colonne G dans la feuille de calcul)
- De la même manière, nous déduisons le résidu $e_y = Y - \hat{Y}$ après la régression $\hat{Y} = 0.01093Z_2 + 0.04434Z_2 - 15.58838$ (colonne H dans la feuille EXCEL)
- Il ne nous reste plus qu'à calculer la corrélation entre les résidus pour obtenir la corrélation partielle d'ordre 2, relativement à Z_1 et Z_2, $\hat{r} = 0.25309$.
- Exactement la même valeur qu'avec l'approche récursive.

Avec les logiciels d'économétrie usuels, nulle doute que cette seconde approche est quand même très facile à mettre en oeuvre, les risques de mauvaises manipulations sont réduits.

	A	B	C	D	E	F	G	H
1			X	Y	Z1	Z2	Résidus	
2	Numero	Modele	Conso	Puissance	Cylindree	Poids	e_X	e_Y
3	1	Daihatsu Cuore	5.7	32	846	650	0.30185	2.96973
4	2	Suzuki Swift 1.0 GLS	5.8	39	993	790	-0.40963	1.92116
5	3	Fiat Panda Mambo L	6.1	29	899	730	0.27854	-3.25479
6	4	VW Polo 1.4 60	6.5	44	1390	955	-0.95753	-12.48664
7	5	Opel Corsa 1.2i Eco	6.8	33	1195	895	-0.13777	-14.18384
8	6	Subaru Vivio 4WD	6.8	32	658	740	1.24824	10.32283
9	7	Toyota Corolla	7.1	55	1331	1010	-0.52422	0.52858
10	8	Opel Astra 1.6i 16V	7.4	74	1597	1080	-0.88075	6.96807
11	9	Peugeot 306 XS 108	9.0	74	1761	1100	0.41702	-0.52292
12	10	Renault Safrane 2.2. V	11.7	101	2165	1500	0.81935	4.19065
13	11	Seat Ibiza 2.0 GTI	9.5	85	1983	1075	0.73849	0.90593
14	12	VW Golt 2.0 GTI	9.5	85	1984	1155	0.38293	-0.01271
15	13	Citroen ZX Volcane	8.8	89	1998	1140	-0.26889	3.53041
16	14	Fiat Tempra 1.6 Liberty	9.3	65	1580	1080	1.04140	-1.27808
17	15	Fort Escort 1.4i PT	8.6	54	1390	1110	0.45609	-4.18058
18	16	Honda Civic Joker 1.4	7.7	66	1396	1140	-0.58458	7.22550
19	17	Volvo 850 2.5	10.8	106	2435	1370	0.14326	-1.36151
20	18	Ford Fiesta 1.2 Zetec	6.6	55	1242	940	-0.59828	5.24020
21	19	Hyundai Sonata 3000	11.7	107	2972	1400	0.21079	-24.50210
22	20	Lancia K 3.0 LS	11.9	150	2958	1550	-0.23522	17.47942
23	21	Mazda Hachtback V	10.8	122	2497	1330	0.23962	12.32631
24	22	Mitsubishi Galant	7.6	66	1998	1300	-2.17742	-21.21818
25	23	Opel Omega 2.5i V6	11.3	125	2496	1670	-0.76470	11.65491
26	24	Peugeot 806 2.0	10.8	89	1998	1560	-0.12878	-1.05962
27	25	Nissan Primera 2.0	9.2	92	1997	1240	-0.31042	5.48189
28	26	Seat Alhambra 2.0	11.6	85	1984	1635	0.35734	-5.25845
29	27	Toyota Previa salon	12.8	97	2438	1800	0.23518	-15.19386
30	28	Volvo 960 Kombi aut	12.7	125	2473	1570	1.10809	13.76769
31								
32								
33				Reg X/Z1,Z2			r	0.25309
34				Z2	Z1	Const.		
35				0.00443	0.00130	1.41755		
36								
37				Reg Y/Z1,Z2				
38				Z2	Z1	Const.		
39				0.01093	0.04434	-15.58838		

Fig. 4.4. Corrélation partielle d'ordre 2 - Approche résidus de régressions - Fichier "voitures"

4.3.2 Test de significativité et intervalle de confiance

Pour tester la significativité avec le t de Student, et calculer les intervalles de confiance à travers la transformation de Fisher, il nous faut généraliser à p variables de contrôle les indicateurs développés dans la section précédente. La principale modification va porter sur l'évaluation des degrés de liberté[3].

Ainsi, la statistique du test de significativité s'écrit maintenant

$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-p-2}}} \tag{4.8}$$

Elle suit une loi de Student à $(n-p-2)$ degrés de liberté.

Et la variance de la transformation de Fisher calculée sur la corrélation partielle d'ordre p devient

$$\sigma_f^2 = \frac{1}{n-p-2} \tag{4.9}$$

3. Voir `http://www.stat.psu.edu/online/development/stat505/07_partcor/06_partcor_partial.html`

4.3.3 Exemple

Finissons notre exemple de corrélation partielle d'ordre 2 sur le fichier voitures (Figure 4.3). Nous pouvons détailler maintenant le contenu des H et I de la feuille EXCEL.

Concernant le test de significativité :
- La corrélation partielle est $\hat{r} = 0.25309$
- Nous calculons t à l'aide de l'équation 4.8, $t = \frac{0.25\hat{3}09}{\sqrt{\frac{1-0.25309^2}{28-2-2}}} = 1.28161$
- Le seuil critique au risque 5% pour un test bilatéral est $t_{0.975}(28-2-2) = 2.39095$. Les données sont compatibles avec l'absence de lien entre "puissance" et "consommation", une fois retranchée l'information apportée par "cylindrée" et "poids".

Concernant l'intervalle de confiance au niveau 95% :
- Nous appliquons tout d'abord la transformation de Fisher : $f = \frac{1}{2}\ln\frac{1+0.25309}{1-0.25309} = 0.25871$
- L'écart type estimé est $\sigma_f = \sqrt{\frac{1}{28-2-2}} = 0.20412$
- Le quantile d'ordre 975% est $u_{0.975} = 1.95996$
- La borne basse (resp. haute) pour f est $bb_f = 0.25871 - 1.95996 \times 0.20412 = -0.14136$ (resp. $bh_f = 0.25871 + 1.95996 \times 0.20412 = 0.69661$)
- Il ne reste plus qu'à appliquer la transformation inverse pour obtenir la borne basse (resp. haute) du coefficient $bb_r = \frac{e^{2\times(-0.14136)}-1}{e^{2\times(-0.14136)}+1} = -0.14043$ (resp. $bh_r = 0.57756$).
- Le résultat est cohérent avec le test d'hypothèses, l'intervalle de confiance englobe la valeur 0.

4.4 Corrélation partielle sur les rangs - ρ de Spearman partiel

Lorsque la relation est non linéaire, le coefficient de Pearson détecte traduit mal l'intensité de la liaison. C'est ce qui avait motivé la présentation du coefficient de Spearman ci-dessus, qui est un coefficient de Pearson calculé sur les rangs. Son avantage est d'être non paramétrique, il permet aussi de mieux rendre compte de la liaison tant qu'elle est monotone. Est-ce que cette approche reste d'actualité concernant la corrélation partielle ?

La réponse est oui. Nous pouvons nous appuyer sur les 2 dispositifs décrits pour le coefficient de corrélation de Pearson.

4.4.1 ρ partiels via les résidus de la régression

Pour calculer le coefficient de Spearman partiel d'ordre p sur un échantillon de données $(\hat{\rho}_{xy.z_1\ldots z_p})$, il suffit d'adopter la démarche suivante [4] :

1. Transformer toutes les variables en rangs. Adopter les rangs moyens en cas d'ex-aequo.

4. Voir la documentation en ligne SAS - `http://support.sas.com/documentation/cdl/en/procstat/59629/HTML/default/procstat_corr_sect017.htm`

2. Calculer le résidu ϵ_x (resp. ϵ_y) de la régression des rangs de X (resp. rangs de Y) avec les rangs des variables de contrôle.
3. Le ρ partiel est tout simplement le coefficient de corrélation de Pearson appliqué sur ces 2 résidus c.-à-d.

$$\hat{\rho}_{xy.z_1 \ldots z_p} = \hat{r}_{\epsilon_x \epsilon_y}$$

4. Le dispositif inférentiel reste inchangé, on doit tenir compte de p dans le calcul des degrés de liberté.

4.4.2 ρ partiels via les formules de récurrence

De la même manière que pour le coefficient de Pearson, nous pouvons utiliser les formules de récurrence (équations 4.1, 4.6 et 4.5) pour calculer les ρ de Spearman partiels de proche en proche. Cette technique est plus simple tant que p est faible (de l'ordre de 1 ou 2 maximum).

4.4.3 Exemple : corrélation entre 2 types de cancer en contrôlant l'effet de la cigarette

Hé ben non, ce n'est pas un exemple sur les voitures ! On cherche à déterminer sur cet exemple s'il existe une part non expliquée par la consommation de cigarettes dans la relation entre l'occurrence du cancer du poumon et celui du cancer de la vessie. Les individus sont des états des USA, CIG (Z) est le nombre de cigarettes par tête fumées, BLAD (X) est le nombre de personnes mortes du cancer de la vessie par 100.000 habitants, et LUNG est le nombre de personnes mortes du cancer de la vessie par 100.000 habitants [5]. La corrélation brute entre BLAD et LUNG est de $\hat{r}_{xy} = 0.6251$, assez forte. Essayons de relativiser cela en contrôlant le rôle de la cigarette.

Décrivons l'organisation de la feuille de calcul (Figure 4.5).

- Les variables sont transformées en rangs, nous créons les variables R, S et T à partir de X, Y et Z. Attention, en cas d'ex-aequo, nous utilisons les rangs moyens.
- Nous disposons de $n = 42$ observations.
- La corrélation brute entre X et Y est $\hat{\rho}_{xy} = 0.6251$.
- Les corrélations brutes avec la variable de contrôle sont $\hat{\rho}_{xz} = 0.6213$ et $\hat{\rho}_{yz} = 0.7264$.
- Nous appliquons la formule 4.2 pour obtenir

$$\hat{\rho}_{xy.z} = \frac{0.6251 - 0.6213 \times 0.7264}{\sqrt{1 - 0.6213^2} \times \sqrt{1 - 0.7264^2}} = 0.32280$$

- Le t de Student associé est

$$t = \frac{0.32280}{\sqrt{\frac{1-0.32280^2}{42-1-2}}}$$

- Avec la loi de Student à $(n - 1 - 2 = 39)$ degrés de liberté, nous obtenons une p-value de 0.0395
- Au risque 5%, on rejette l'hypothèse nulle. Il semble qu'il y ait autre chose non expliquée par la cigarette dans la liaison entre les 2 types de cancer (ceci étant à 1% la liaison n'est pas significative, la liaison partielle est assez tenue).

5. `http://lib.stat.cmu.edu/DASL/Stories/cigcancer.html` - Nous avons supprimé du fichier les 2 états signalés atypiques.

	A	B	C	D	E	F	G
1		Z	X	Y			
2	STATE	CIG	BLAD	LUNG	R	S	T
3	UT	14	3.31	12.01	1	14	1
4	MS	16.08	3.06	15.6	2	9	9
5	SC	18.06	3.25	17.45	3	13	16
6	AL	18.2	2.9	17.05	4	3.5	15
7	AR	18.24	2.99	15.98	5	8	11
8	ND	19.96	2.89	12.12	6	2	3
9	TE	20.08	2.94	17.6	7	7	17
10	ID	20.1	3.08	13.58	8	10	4
11	SD	20.94	3.64	14.11	9	17	5
12	NM	21.16	2.9	14.59	10	3.5	7
13	WA	21.17	4.04	20.34	11	21.5	23
14	WI	21.25	5.14	20.55	12	38	24
15	LA	21.58	4.65	25.45	13	29	39
16	KS	21.84	2.91	16.84	14	5	14
17	MN	22.06	3.72	14.2	15	18.5	6
18	IO	22.12	4.23	16.59	16	24	12
19	TX	22.57	3.21	20.74	17	12	25
20	WV	22.86	4.78	15.53	18	32.5	8
21	NB	23.32	3.72	16.7	19	18.5	13
22	KY	23.44	2.86	17.71	20.5	1	18
23	OK	23.44	2.93	19.45	20.5	6	19
24	MT	23.75	3.95	19.5	22	20	20
25	PE	23.78	4.89	12.11	23	35	2
26	MI	24.96	5.27	22.72	24	40	32
27	AZ	25.82	3.52	19.8	25	16	21
28	VT	25.89	4.63	21.22	26	28	28
29	MD	25.91	5.21	26.48	27	39	42
30	IN	26.18	4.09	20.3	28	23	22
31	OH	26.38	4.47	21.89	29	27	29
32	MA	26.92	4.69	22.04	30	30	30
33	MO	27.56	4.04	20.98	31	21.5	27
34	IL	27.91	4.75	22.8	32	31	33
35	WY	28.04	3.2	15.92	33	11	10
36	FL	28.27	4.46	23.57	34	25.5	35
37	CA	28.6	4.46	22.07	35	25.5	31
38	NJ	28.64	5.98	25.95	36	42	41
39	ME	28.92	4.79	20.94	37	34	26
40	NY	29.14	5.3	25.02	38	41	38
41	RI	29.18	4.99	23.68	39	36	36
42	AK	30.34	3.46	25.88	40	15	40
43	CT	31.1	5.11	22.83	41	37	34
44	DE	33.6	4.78	24.55	42	32.5	37
45							
46						n	42
47			Rho de Spearman				
48		X	Y	0.6251		rho (partiel)	0.32280
49		X	Z	0.6213			
50		Y	Z	0.7264		t	2.12988
51						p-value	0.0395

Fig. 4.5. ρ de Spearman partiel d'ordre 1 - Approche récursive

A titre de comparaison, voici les commandes et sorties SAS (Figure 4.6). Les résultats concordent. C'est préférable étant donné qu'on a suivi à la lettre le descriptif de la documentation en ligne.

Remarque 15 (Corrélation partielle basée sur le τ de Kendall). Il est possible de calculer le τ partiel de Kendall à partir des τ bruts en utilisant la formule de passage analogue à celle du coefficient de Pearson (équation 4.1) (voir [9], page 254 à 262 ; ou son résumé en français sur le site `http://www.cons-dev.org/elearning/stat/stat7/st7.html`). On peut très bien la mettre en oeuvre lorsque les données sont intrinsèquement des classements (des rangs affectés). Malheureusement, les avis divergent quant au calcul de la distribution de la statistique, le test de significativité est difficile, ce qui est un frein considérable à son utilisation.

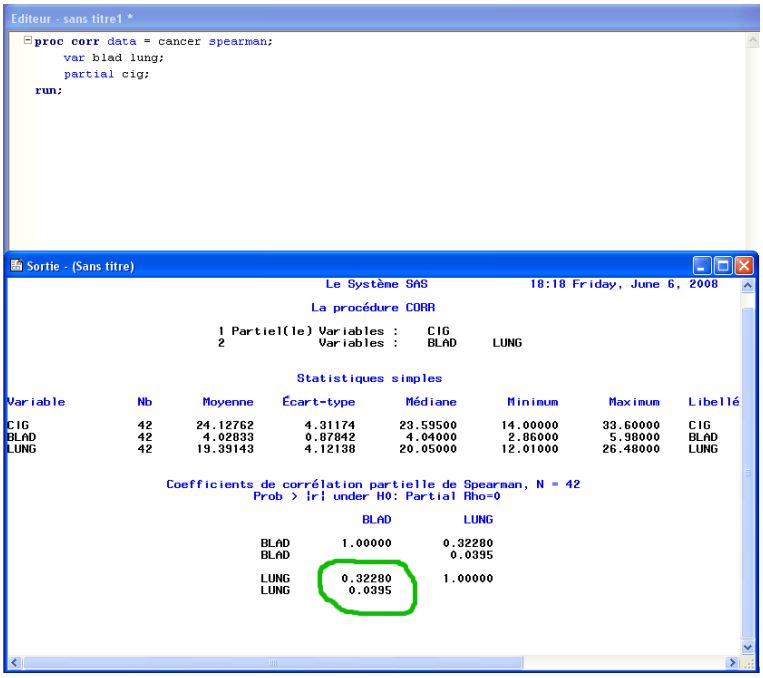


Fig. 4.6. ρ de Spearman partiel d'ordre 1 - Commandes et sorties SAS

5

Corrélation semi-partielle

5.1 Principe de la corrélation semi-partielle

A la différence de la corrélation partielle, la corrélation semi-partielle[1] est résolument asymétrique, elle se rapproche de la régression multiple. On essaie de quantifier le pouvoir explicatif additionnel d'une variable.

Positionnons nous dans un premier temps dans le cadre à 3 variables Y, X, et Z : Y est la variable dépendante que l'on cherche à expliquer, X est la variable explicative que l'on cherche à évaluer, Z est la variable de contrôle. Le carré de la corrélation semi-partielle, notée $r^2_{y(x.z)}$, quantifie la proportion de variance de Y expliquée par X, sachant que l'on a retranché de cette dernière l'information apportée par Z. En d'autres termes, quelle est la part de Y qu'explique l'information additionnelle de X par rapport à Z.

Notons bien la différence avec la corrélation partielle. Avec ce dernier, nous retranchons l'information apportée par Z sur à la fois Y et X, et nous mesurons la liaison sur les résidus. Dans le cas de la corrélation semi-partielle, nous cherchons à quantifier la liaison de Y avec la partie résiduelle de X par rapport à Z. On discerne bien le caractère asymétrique de l'approche.

Dans notre exemple des véhicules (Figure 2.2), nous posons la question suivante : si on enlève de la puissance (X) l'information portée par la cylindrée (Z), est-ce qu'il reste quelque chose pour expliquer la consommation (Y) ? En d'autres termes, on cherche à évaluer l'apport additionnel de puissance (X), par rapport à la cylindrée (Z), dans l'explication de la consommation (Y).

5.2 Calcul et inférence statistique

La corrélation semi-partielle de Y avec X conditionnellement à Z est définie de la manière suivante

$$r_{y(x.z)} = \frac{r_{yx} - r_{yz}r_{xz}}{\sqrt{1 - r^2_{xz}}} \tag{5.1}$$

1. *semi-partial correlation* ou *part correlation* en anglais

Notons d'ores et déjà que $r_{y(x.z)} = r_{yx}$ si X et Z sont orthogonaux $r_{xz} = 0$. Tout l'information de X peut être utilisée pour expliquer Y. Si X et Z sont parfaitement corrélés c.-à-d. $r_{xz} = 1$, l'équation 5.1 est indéfinie, mais on comprend aisément qu'il ne reste plus rien dans le résidu de X pour expliquer Y.

En faisant le parallèle avec la formule de la corrélation partielle (équation 4.1), on constate de manière générale que

$$r_{yx.z} \geq r_{y(x.z)}$$

Estimation. Sur un échantillon de taille n, pour estimer la corrélation semi-partielle, il suffit de remplacer les corrélation théoriques de la formule 5.1 par les corrélations empiriques.

Test de significativité. Pour tester la significativité de la corrélation i.e. $H_0 : r_{y(x.z)} = 0$ (test unilatéral ou bilatéral), nous utilisons le t de Student qui est a la même expression que celle de la corrélation partielle, avec la même distribution et les mêmes degrés de liberté $(n-3)$, à savoir

$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-3}}} \tag{5.2}$$

Exemple : utiliser l'information résiduelle de la puissance (relativement à la cylindrée) pour expliquer la consommation. Reprenons notre fameux fichier des voitures, réalisons les calculs (Figure 5.1) :

	A	B	C	D	E
1			X	Y	Z
2	Numero	Modele	Puissance	Conso	Cylindree
3	1	Daihatsu Cuore	32	5.7	846
4	2	Suzuki Swift 1.0 GLS	39	5.8	993
5	3	Fiat Panda Mambo L	29	6.1	899
6	4	VW Polo 1.4 60	44	6.5	1390
7	5	Opel Corsa 1.2i Eco	33	6.8	1195
8	6	Subaru Vivio 4WD	32	6.8	658
9	7	Toyota Corolla	55	7.1	1331
10	8	Opel Astra 1.6i 16V	74	7.4	1597
11	9	Peugeot 306 XS 108	74	9.0	1761
12	10	Renault Safrane 2.2. V	101	11.7	2165
13	11	Seat Ibiza 2.0 GTI	85	9.5	1983
14	12	VW Golt 2.0 GTI	85	9.5	1984
15	13	Citroen ZX Volcane	89	8.8	1998
16	14	Fiat Tempra 1.6 Liberty	65	9.3	1580
17	15	Fort Escort 1.4i PT	54	8.6	1390
18	16	Honda Civic Joker 1.4	66	7.7	1396
19	17	Volvo 850 2.5	106	10.8	2435
20	18	Ford Fiesta 1.2 Zetec	55	6.6	1242
21	19	Hyundai Sonata 3000	107	11.7	2972
22	20	Lancia K 3.0 LS	150	11.9	2958
23	21	Mazda Hachtback V	122	10.8	2497
24	22	Mitsubishi Galant	66	7.6	1998
25	23	Opel Omega 2.5i V6	125	11.3	2496
26	24	Peugeot 806 2.0	89	10.8	1998
27	25	Nissan Primera 2.0	92	9.2	1997
28	26	Seat Alhambra 2.0	85	11.6	1984
29	27	Toyota Previa salon	97	12.8	2438
30	28	Volvo 960 Kombi aut	125	12.7	2473

	G	H
	n	28

Corrélations brutes		
Puissance	Conso	0.88781
Puissance	Cylindrée	0.94755
Conso	Cylindrée	0.89187

Corrélation semi-partielle	
r_y(x.z)	0.13367

Test de significativité	
t	0.67439
t(0.975 ; 25)	2.38461
p-value	0.50625

Fig. 5.1. Coefficient semi-partiel - Exemple des voitures

- Nous avons $n = 28$
- La corrélation brute entre Y et X est $\hat{r}_{yx} = 0.88781$, la liaison semble forte.
- Les autres corrélations brutes sont $\hat{r}_{xz} = 0.94755$ et $\hat{r}_{yz} = 0.89187$

– Nous formons l'équation 5.1

$$\hat{r}_{y(x.z)} = \frac{0.88781 - 0.89187 \times 0.94755}{\sqrt{(1 - 0.94755^2}} = 0.13367$$

– le t de Student pour le test de significativité est

$$t = \frac{0.13367}{\sqrt{\frac{1-0.13367^2}{28-3}}} = 0.67439$$

– Au risque 5%, le seuil critique est $t_{0.975}(25) = 2.38461$. Nous acceptons l'hypothèse de nullité du coefficient. Manifestement, une fois retranchée de "puissance" l'information portée par "cylindrée", il ne reste plus rien pour expliquer la "consommation".

5.3 Corrélation semi-partielle d'ordre p

Il est possible de généraliser la notion de corrélation semi-partielle à p variables de contrôle. Il s'agit de calculer la liaison entre Y et X, une fois retranchée de cette dernière l'influence de $Z_1 \dots Z_p$ variables. Pour réaliser le calcul pratique du coefficient, nous utilisons la régression, ça nous permet de comprendre autrement, de manière plus générique, le mécanisme d'évaluation de la liaison.

Concernant l'inférence statistique, le **test de significativité** est très similaire à la corrélation partielle, notamment en ce qui concerne le calcul des degrés de liberté. Pour tester la significativité, nous utiliserons la statistique t qui, sous l'hypothèse de nullité du coefficient, suit une loi de Student à $(n - p - 2)$ degrés de liberté

$$t = \frac{\hat{r}}{\sqrt{\frac{1-\hat{r}^2}{n-p-2}}} \tag{5.3}$$

5.3.1 Utilisation des résidus de la régression

Une bonne manière de construire la corrélation partielle est de prendre au pied la lettre la définition en utilisant les résidus de la régression. Voici la séquence des traitements :

– Dans un premier temps, nous calculons la régression linéaire multiple

$$X = a_0 + a_1 Z_1 + \dots a_p Z_p + \epsilon$$

– A partir des coefficients estimés $\hat{a}_j$, nous déduisons les valeurs prédites $\hat{X}$

– Nous construisons alors les résidus de la régression qui représente la fraction de X (l'information que porte X) qui n'est pas déjà expliquée par les variables de contrôle.

$$e_i = x_i - \hat{x}_i$$

– La corrélation semi partielle estimée est obtenue à l'aide de la corrélation empirique entre Y et le résidu e

$$\hat{r}_{y(x.z_1 \cdots z_p)} = \hat{r}_{ye} \tag{5.4}$$

5.3.2 Comparaison de régressions

Une approche alternative pour calculer la corrélation semi-partielle est de comparer différentes régressions expliquant Y. En effet, on cherche à quantifier le pouvoir explicatif additionnel de X par rapport aux variables de contrôle. Le carré du coefficient s'interprète lui-même comme une proportion de variance expliquée supplémentaire. A partir de ce point de vue, on peut proposer une autre manière d'estimer le coefficient de corrélation semi-partielle. Voici la séquence de calculs :

- On effectue une première régression de Y sur les variables de contrôle $Z_1, \ldots, Z_p$, nous obtenons le coefficient de détermination $R^2_{y.z_1 \cdots z_p}$, il correspond à la proportion de variance expliquée par la régression.
- On réaliser une seconde régression intégrant la variable supplémentaire X parmi les explicatives, un nouveau coefficient de détermination $R^2_{u.xz_1 \cdots z_p}$ est dégagé.
- Le surcroît d'information qu'apporte X dans l'explication de Y, par rapport aux variables de contrôle, est la différence entre les R^2. C'est aussi le carré du coefficient de corrélation semi-partielle

$$\hat{r}^2_{y(x.z_1 \cdots z_p)} = R^2_{y.xz_1 \cdots z_p} - R^2_{y.z_1 \cdots z_p} \tag{5.5}$$

- La racine carrée de cette quantité est le résultat souhaité.

5.3.3 Exemple d'application

La démarche est générique pour ($p \geq 1$). Néanmoins, pour illustrer notre propos, nous reprenons notre exemple de la section consacrée à la corrélation semi-partielle d'ordre 1 (section 5.2). L'intérêt est de pouvoir comparer les coefficients obtenus selon les différents approches. Les calculs sont regroupés dans une nouvelle feuille (figure 5.2).

Détaillons tout d'abord l'approche basée sur la comparaison de régressions :

- La régression de Y sur la variable de contrôle Z fournit $R^2_{y.z} = 0.79543$. Nous avons utilisé la fonction DROITEREG() d'EXCEL.
- La régression de Y sur X et Z fournit $R^2_{y.xz} = 0.81329$
- Le gain d'explication consécutif à l'introduction de X dans la régression est donc $\Delta = 0.81329 - 0.7953 = 0.01787$
- Et sa racine carrée est la corrélation semi-partielle $\hat{r}_{y(x.z)} = \sqrt{0.01787} = 0.13367$. Nous obtenons exactement la même valeur qu'avec la méthode directe décrite dans la section 5.2.

Détaillons maintenant l'approche basée sur les résidus de la régression :

- Nous réalisons la régression de X sur la variable de contrôle Z. Nous utilisons les coefficients pour calculer la colonne des résidus qui correspond à la fraction de X non expliquée par Z

$$e_i = x_i - (0.04901 \times z_i - 10.94646)$$

- Nous calculons la corrélation de Pearson entre le résidu e et la variable Y, elle correspond à la corrélation semi-partielle $\hat{r}_{y(x.z)} = \hat{r}_{ye} = 0.13367$. De nouveau la valeur obtenue est cohérente avec celles proposées par les approches alternatives.

	A	B	C	D	E	F	G	H	I	J
1			Y	X	Z					
2	Numero	Modele	Conso	Puissance	Cylindree	e				
3	1	Daihatsu Cuore	5.7	32	846	1.485				
4	2	Suzuki Swift 1.0 GLS	5.8	39	993	1.281				
5	3	Fiat Panda Mambo L	6.1	29	899	-4.113				
6	4	VW Polo 1.4 60	6.5	44	1390	-13.176				
7	5	Opel Corsa 1.2i Eco	6.8	33	1195	-14.619				
8	6	Subaru Vivio 4WD	6.8	32	658	10.699				
9	7	Toyota Corolla	7.1	55	1331	0.716				
10	8	Opel Astra 1.6i 16V	7.4	74	1597	6.679				
11	9	Peugeot 306 XS 108	9.0	74	1761	-1.358				
12	10	Renault Safrane 2.2. V	11.7	101	2165	5.842				
13	11	Seat Ibiza 2.0 GTI	9.5	85	1983	-1.238				
14	12	VW Golt 2.0 GTI	9.5	85	1984	-1.287				
15	13	Citroen ZX Volcane	8.8	89	1998	2.027				
16	14	Fiat Tempra 1.6 Liberty	9.3	65	1580	-1.488				
17	15	Fort Escort 1.4i PT	8.6	54	1390	-3.176				
18	16	Honda Civic Joker 1.4	7.7	66	1396	8.530				
19	17	Volvo 850 2.5	10.8	106	2435	-2.390				
20	18	Ford Fiesta 1.2 Zetec	6.6	55	1242	5.077				
21	19	Hyundai Sonata 3000	11.7	107	2972	-27.708				
22	20	Lancia K 3.0 LS	11.9	150	2958	15.978				
23	21	Mazda Hachtback V	10.8	122	2497	10.571				
24	22	Mitsubishi Galant	7.6	66	1998	-20.973				
25	23	Opel Omega 2.5i V6	11.3	125	2496	13.620				
26	24	Peugeot 806 2.0	10.8	89	1998	2.027				
27	25	Nissan Primera 2.0	9.2	92	1997	5.076				
28	26	Seat Alhambra 2.0	11.6	85	1984	-1.287				
29	27	Toyota Previa salon	12.8	97	2438	-11.537				
30	28	Volvo 960 Kombi aut	12.7	125	2473	14.747				

Ecart entre les régression

Régression Y / Z

0.00319	3.29846
0.00053	0.66685
0.79543	1.02916

Régression Y / X,Z

0.00177	0.02895	3.61536
0.00097	0.01872	0.62544
0.81329	1.05285	#N/A

r^2 = DIFF (R^2)	0.01787
r	**0.13367**

Utilisation des résidus de la régression

Régression X /Z

Z	const
0.04901	-10.94646

r	**0.13367**

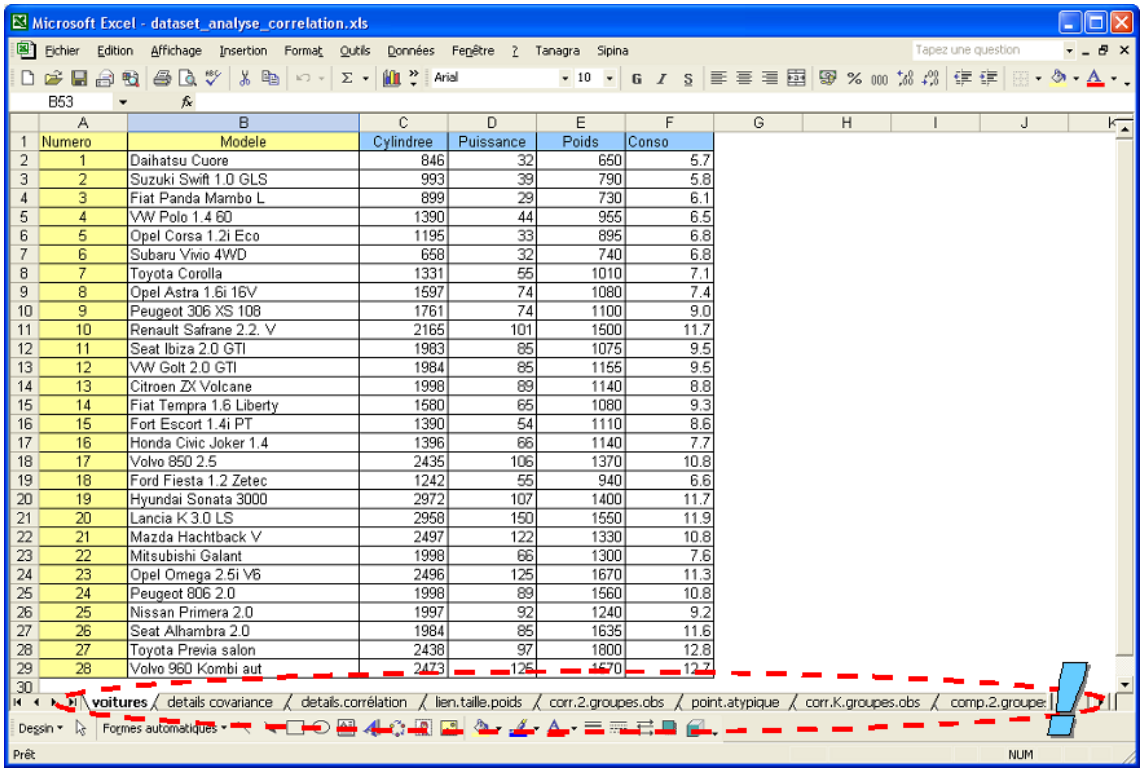
Fig. 5.2. Coefficient semi-partiel - Approche régression - Exemple des voitures

A

Fichier de données

Tout au long de ce support, nous illustrons notre propos à l'aide d'exemples numériques. Les données et les calculs associés sont disponibles dans un classeur EXCEL accessible en ligne. L'URL du fichier est `http://eric.univ-lyon2.fr/~ricco/cours/cours/dataset_analyse_correlation.xls`.

A chaque feuille du classeur correspond un thème du support. Pour faire la correspondance, le plus simple est de se référer à l'onglet de la feuille (Figure A.1).



	A	B	C	D	E	F
1	Numero	Modele	Cylindree	Puissance	Poids	Conso
2	1	Daihatsu Cuore	846	32	650	5.7
3	2	Suzuki Swift 1.0 GLS	993	39	790	5.8
4	3	Fiat Panda Mambo L	899	29	730	6.1
5	4	VW Polo 1.4 60	1390	44	955	6.5
6	5	Opel Corsa 1.2i Eco	1195	33	895	6.8
7	6	Subaru Vivio 4WD	658	32	740	6.8
8	7	Toyota Corolla	1331	55	1010	7.1
9	8	Opel Astra 1.6i 16V	1597	74	1080	7.4
10	9	Peugeot 306 XS 108	1761	74	1100	9.0
11	10	Renault Safrane 2.2. V	2165	101	1500	11.7
12	11	Seat Ibiza 2.0 GTI	1983	85	1075	9.5
13	12	VW Golt 2.0 GTI	1984	85	1155	9.5
14	13	Citroen ZX Volcane	1998	89	1140	8.8
15	14	Fiat Tempra 1.6 Liberty	1580	65	1080	9.3
16	15	Fort Escort 1.4i PT	1390	54	1110	8.6
17	16	Honda Civic Joker 1.4	1396	66	1140	7.7
18	17	Volvo 850 2.5	2435	106	1370	10.8
19	18	Ford Fiesta 1.2 Zetec	1242	55	940	6.6
20	19	Hyundai Sonata 3000	2972	107	1400	11.7
21	20	Lancia K 3.0 LS	2958	150	1550	11.9
22	21	Mazda Hachtback V	2497	122	1330	10.8
23	22	Mitsubishi Galant	1998	66	1300	7.6
24	23	Opel Omega 2.5i V6	2496	125	1670	11.3
25	24	Peugeot 806 2.0	1998	89	1560	10.8
26	25	Nissan Primera 2.0	1997	92	1240	9.2
27	26	Seat Alhambra 2.0	1984	85	1635	11.6
28	27	Toyota Previa salon	2438	97	1800	12.8
29	28	Volvo 960 Kombi aut	2473	125	1670	12.7
30						

Fig. A.1. Classeur EXCEL - Analyse de corrélation

B

L'analyse de corrélation avec Tanagra

Les techniques présentées dans ce support sont implémentés dans le logiciel gratuit et *open source* Tanagra – `http://eric.univ-lyon2.fr/~ricco/tanagra/`.

Leur mise en oeuvre et la lecture des résultats sont décrites dans plusieurs didacticiels, en voici quelques uns :

1. Corrélation semi-partielle
   `http://tutoriels-data-mining.blogspot.com/2008/06/corrlation-semi-partielle.html`
2. Corrélation partielle
   `http://tutoriels-data-mining.blogspot.com/2008/06/corrlation-partielle.html`
3. Corrélations croisées
   `http://tutoriels-data-mining.blogspot.com/2008/04/coefficient-de-corrlation-linaire.html`
4. De manière générale, on pourra accéder aux didacticiels qui abordent le coefficient de corrélation linéaire et ses variantes en effectuant une recherche par mots clés sur le site de tutoriels
   `http://tutoriels-data-mining.blogspot.com/`

Littérature

Ouvrages

1. Aïvazian, S., *Etude statistique des dépendances*, Mir, Moscou, 1978.
2. Chen, P., Popovich, P., *Correlation : Parametric and Nonparametric Measures*, Sage University Papers Series on Quantitative Applications in the Social Sciences, no. 07-139, 2002.
3. Dodge, Y, Rousson, V., *Analyse de régression appliquée*, Dunod, 2004.
4. Johnston, J., DiNardo, J., *Méthodes Econométriques*, Economica, 4è édition, 1999.
5. Garson, D., *Statnotes : Topics in Multivariate Analysis*, `http://www2.chass.ncsu.edu/garson/pa765/statnote.htm`.
6. Howell, D., *Méthodes statistiques en sciences humaines*, De Boeck Université, 1998.
7. Saporta, G., *Probabilités, Analyse de Données et Statistique*, Dunod, 2006.
8. SAS Institute Inc., *SAS 9.1 Documentation*, `http://support.sas.com/documentation/onlinedoc/91pdf/index.html`
9. Siegel, S., Castellan Jr., J., *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Inc., Second Edition, 1988.
10. Veysseyre, R., *Aide-mémoire - Statistique et probabilités pour l'ingénieur*, Dunod, 2006.