Fiche de Biostatistique - Stage 2

La géométrie de l'espace des variables

D. Chessel & A.B. Dufour

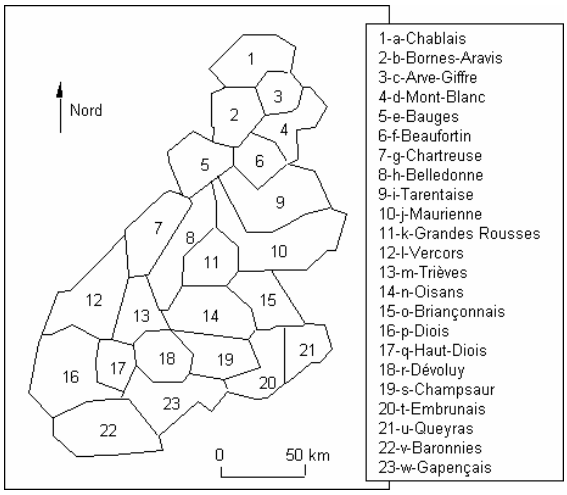
Résumé

La fiche donne les principes de la statistique basée sur une approche géométrique de l'ensemble des variables. Elle contient les démonstrations importantes pour une lecture précise mais on peut s'en servir comme introduction aux tests sur le modèle linéaire.

Plan

1. Introduction

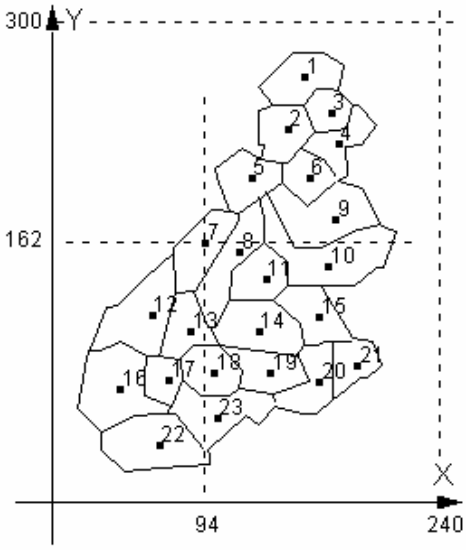
Tableau individus-variables. Définition des lignes-individus et des colonnes-variables.



	A	B	C	D	E	F	G
1	-6.6	1.1	9.7	22.9	131.5	139.5	1597
2	-7.5	0.1	8.2	20.8	115	145	1613
3	-8.5	-0.1	8.6	22	113	146.5	1738
4	-9.2	-1.7	6.5	17.3	103	138	1630
5	-8	0.2	7.5	21.6	110	129	1492
6	-7.8	0.5	8.6	21	113	130	1415
7	-2.9	2.8	11.5	19.8	188.5	141.5	1849
8	-8.4	-1	8.3	21.1	100	120	1473
9	-6.8	2.3	8.8	21.4	100	85	978.5
10	-6.8	0.8	9.5	22.8	75	38	784.5
11	-6.3	1.9	8.1	19.2	77	80.5	976
12	-4	5.5	10.5	24.2	69	79	1239
13	-7.2	1.2	9.6	22.9	65	72.5	1125
14	-10.1	0.2	5.3	19.7	65	70	1025
15	-8.9	2.2	7.6	22.7	65	41	771.5
16	-1.6	7.3	12.6	27.6	51	47.5	920
17	-4	5	10	24.5	66	57.5	1010
18	-6.6	2.8	8.9	24.7	98.5	52	1116
19	-7	1.8	8.6	21.9	112	68	1248
20	-6.3	3.5	10.5	24.3	57.5	55	767.5
21	-8.6	2.6	7.6	22.4	49.5	48.5	766.5
22	-3.4	7.3	12.5	27.4	56.5	29.5	926.5
23	-6.6	4	9.8	26	77.5	47	955

A [minja] - Mini Janvier (°C) B [maxja] - Maxi Janvier (°C) C [minju] - Mini Juillet (°C)
D [maxju] - Maxi Juillet (°C) E [pja] - Pluviométrie Janvier (mm)
F [pju]– Pluviométrie Juillet (mm) G [ptot] – Pluviométrie annuelle (mm)

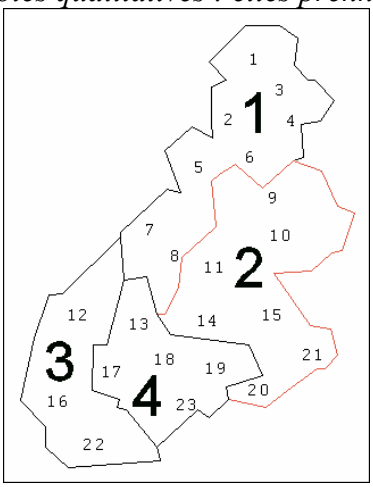
Variables explicatives : celles dont on se sert pour prédire les autres.



	X	Y	Z		X	Y	Z
1	156	252	1685	13	85	97	1040
2	141	217	1470	14	125	99	2480
3	171	233	1680	15	161	103	2410
4	178	215	2500	16	45	62	755
5	123	189	1000	17	69	69	1105
6	154	195	1840	18	97	71	1535
7	94	152	1055	19	129	71	1915
8	110	137	1405	20	154	66	1875
9	167	171	2155	21	187	79	2390
10	167	148	2290	22	62	29	975
11	128	130	2200	23	102	49	970
12	49	101	1030				

X – Longitude Y - Latitude Z – Altitude R - Région

Variables qualitatives : elles prennent des valeurs dans un ensemble de modalités.



	R		R		R		R
1	1	7	1	13	4	19	4
2	1	8	1	14	2	20	2
3	1	9	2	15	2	21	2
4	1	10	2	16	3	22	3
5	1	11	2	17	4	23	4
6	1	12	3	18	4		

Tableau floristique : le code espèce.

a1	Chêne pubescent (Quercus pubescens Wild.)
a2	Charme (Carpinus betulus L.)
a3	Aulne vert (Alnus viridis Chaix)
a4	Mélèze (Larix europaea D.C.)
a5	Pin sylvestre (Pinus sylvestris L.)
a6	Chêne pédonculé (Quercus pedunculata Ehrh.)
a7	Chêne sessile (Quercus sessiliflora Salisb.)
a8	Sapin (Abies pectinata D.C.)
a9	Hêtre (Fagus sylvatica L.)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
a1	1	1	1	0	1	1	1	2	1	1	1	1	2	1	0	3	2	1	1	2	0	3	3
a2	2	1	1	0	2	0	1	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0
a3	1	1	2	3	0	2	1	1	3	2	2	1	1	2	1	0	0	0	2	1	2	0	0
a4	0	0	1	1	0	1	0	0	2	2	2	0	1	2	3	0	0	1	2	3	3	0	1
a5	0	1	0	1	0	0	1	1	2	2	1	1	2	1	2	3	3	3	2	3	2	3	3
a6	1	1	1	1	1	0	1	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
a7	2	1	1	1	2	1	1	2	2	1	1	1	2	1	0	1	0	0	0	0	0	0	0
a8	2	2	2	2	3	2	3	2	2	2	2	3	2	1	1	1	3	2	2	1	1	1	0
a9	3	3	2	1	3	2	3	2	1	1	1	0	2	1	0	2	0	2	2	0	0	2	2

Tableau faunistique :

	Fauvette orphée	Fauvette des jardins	Fauvette à tête noire	Fauvette babillarde	Fauvette grisette	Fauvette pitchou	Fauvette passerinette	Fauvette mélanocéphale	Hibou petit duc	Chouette de Tengmalm	Chouette chevêche	Chouette chevêchette	Chouette hulotte	Chouette effraie	Bruant fou	Bruant ortolan	Bruant zizi	Bruant jaune	Bruant proyer
1	0	2	2	1	1	0	0	0	0	2	0	1	2	0	2	0	0	2	0
2	0	2	2	1	1	0	0	0	0	2	1	1	2	1	2	0	1	2	0
3	0	2	2	2	1	0	0	0	0	2	1	1	3	1	2	0	1	2	0
4	0	3	2	1	0	0	0	0	0	1	1	1	2	1	1	0	1	2	1
5	0	3	2	1	2	0	0	0	0	1	1	1	2	1	1	1	1	2	1
6	0	2	2	1	0	0	0	0	0	1	1	0	2	1	1	0	1	2	0
7	0	2	3	1	1	0	0	0	0	2	2	1	3	2	1	1	1	2	1
8	0	2	2	1	2	0	0	0	0	1	1	0	2	1	1	0	1	2	0
9	0	3	2	2	1	0	0	0	1	2	1	0	2	1	2	2	1	2	0
10	1	3	2	3	1	0	0	0	1	2	1	1	2	1	3	2	1	3	1
11	0	3	2	2	1	0	0	0	0	1	0	0	1	0	1	1	1	3	0
12	1	2	2	2	2	0	0	0	1	2	1	1	3	2	2	1	1	3	0
13	0	2	2	1	1	0	0	0	0	0	1	0	2	1	1	2	1	3	1
14	1	3	3	2	1	0	0	0	1	1	1	1	2	1	2	1	1	1	0
15	1	2	2	1	2	0	0	0	1	1	0	1	1	1	3	2	2	3	1
16	2	2	3	1	3	1	3	0	3	1	2	0	3	2	2	3	3	1	2
17	1	3	2	0	1	0	0	0	1	1	1	0	3	1	2	2	1	3	1
18	0	2	1	2	2	0	0	0	0	1	1	0	3	0	3	3	2	3	1
19	0	3	2	2	2	0	0	0	0	2	2	0	3	1	3	3	2	3	1
20	1	3	3	2	2	0	0	0	1	2	1	1	3	1	3	3	2	3	1
21	0	2	2	2	2	0	0	0	0	2	1	1	2	0	2	2	1	2	0
22	2	2	3	0	3	2	3	2	3	0	2	0	3	2	2	3	3	1	2
23	1	3	3	1	3	0	2	0	2	1	2	0	3	2	3	3	2	1	3

Source : Lebreton, Ph. . (1977) *Les oiseaux nicheurs rhônalpins*. Atlas ornithologique Rhône-Alpes. Centre Ornithologique Rhône-Alpes, Université Lyon 1, 69621 Villeurbanne. Direction de la Protection de la Nature, Ministère de la Qualité de la Vie. 1-354.

Comment mesurer la liaison entre deux variables quantitatives, entre une variable quantitative et une partition, entre une variable quantitative et une liste d'espèces, entre *K* variables quantitatives, entre une variable quantitative et *K* variables quantitatives, entre *K* variables quantitatives et une partition, entre deux partitions … La géométrie euclidienne donne une solution globale à laquelle introduit ce cours. Penser qu'une base de données écologiques peut contenir des dizaines de variables, des centaines d'espèces, des milliers de relevés. Dans tout ce qui suit n est le nombre d'individus statistiques (exemple $n = 23$ districts). Une variable quantitative est un vecteur de $\mathbb{R}^n$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \mathbf{x}^t = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

Un tableau **X** est une collection de *p* variables et de *n* individus :

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}^1 & \mathbf{x}^2 & \cdots & \mathbf{x}^p \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}$$

$\mathbf{1}_n$ est la variable constante, $\mathbf{U}_{np} = \begin{bmatrix} \mathbf{1}_n & \mathbf{1}_n & \cdots & \mathbf{1}_n \end{bmatrix}$ est le tableau constant, $\mathbf{I}_n$ est la matrice identité à *n* lignes et *n* colonnes :

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = Diag\left(\mathbf{1}_n\right)$$

Dans une étude donnée, les individus statistiques ont des poids qui forment le vecteur des poids $\mathbf{p} = \left(p_1, p_2, ..., p_n\right)$. Par définition les poids sont strictement positifs et leur somme vaut l'unité $\sum_{i=1}^{n} p_i = 1$. La pondération la plus simple est la pondération uniforme, soit $p_i = 1/n$. **D** est la diagonale des poids :

$$\mathbf{D} = \begin{bmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_n \end{bmatrix} = Diag\left(\mathbf{p}\right)$$

L'application définie sur $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ par :

$$\left.\begin{aligned} \mathbf{x}^t &= \left[x_1, x_2, ..., x_n\right] \\ \mathbf{y}^t &= \left[y_1, y_2, ..., y_n\right] \end{aligned}\right\} \left\langle \mathbf{x} \middle| \mathbf{y} \right\rangle_{\mathbf{D}} = \sum_{i=1}^{n} p_i x_i y_i = \mathbf{x}^t \mathbf{D} \mathbf{y}$$

est un produit scalaire. On l'appellera le produit scalaire associé à la pondération **p**.

Si les poids des individus statistiques ne sont pas indiqués explicitement, on utilise la pondération uniforme, donc :

$$\mathbf{D} = \left(1/n\right)\mathbf{I}_n$$

$$\left.\begin{aligned} \mathbf{x}^t &= \left[x_1, x_2, ..., x_n\right] \\ \mathbf{y}^t &= \left[y_1, y_2, ..., y_n\right] \end{aligned}\right\} \left\langle \mathbf{x} \middle| \mathbf{y} \right\rangle_{\mathbf{D}} = \sum_{i=1}^{n} \frac{1}{n} x_i y_i = \frac{1}{n} \mathbf{x}^t \mathbf{y}$$

La notion de produit scalaire sert de fondement.

2. Couples de variables quantitatives

2.1. Moyennes et variances

On utilise le produit scalaire associé à la pondération **p**. La moyenne est un produit scalaire :

$$\overline{\mathbf{x}} = m(\mathbf{x}) = \sum_{i=1}^{n} p_i x_i = \langle \mathbf{x} | \mathbf{1}_n \rangle$$

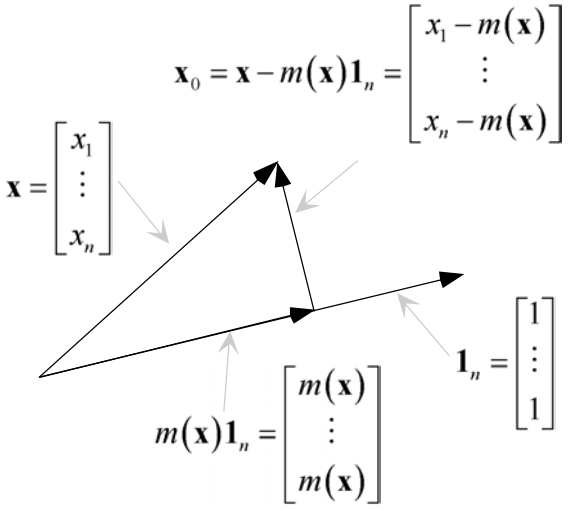
La moyenne des carrés est un produit scalaire :

$$m_2(\mathbf{x}) = \sum_{i=1}^{n} p_i x_i^2 = \langle \mathbf{x} | \mathbf{x} \rangle = \|\mathbf{x}\|^2$$

La variance est le carré d'une norme :

$$v = v(\mathbf{x}) = \sum_{i=1}^{n} p_i \left(x_i - \overline{\mathbf{x}}\right)^2 = \left\|\mathbf{x} - m(\mathbf{x})\mathbf{1}_n\right\|^2$$

$$\mathbf{x}_0 = \mathbf{x} - m(\mathbf{x})\mathbf{1}_n = \begin{bmatrix} x_1 - m(\mathbf{x}) \\ \vdots \\ x_n - m(\mathbf{x}) \end{bmatrix}$$



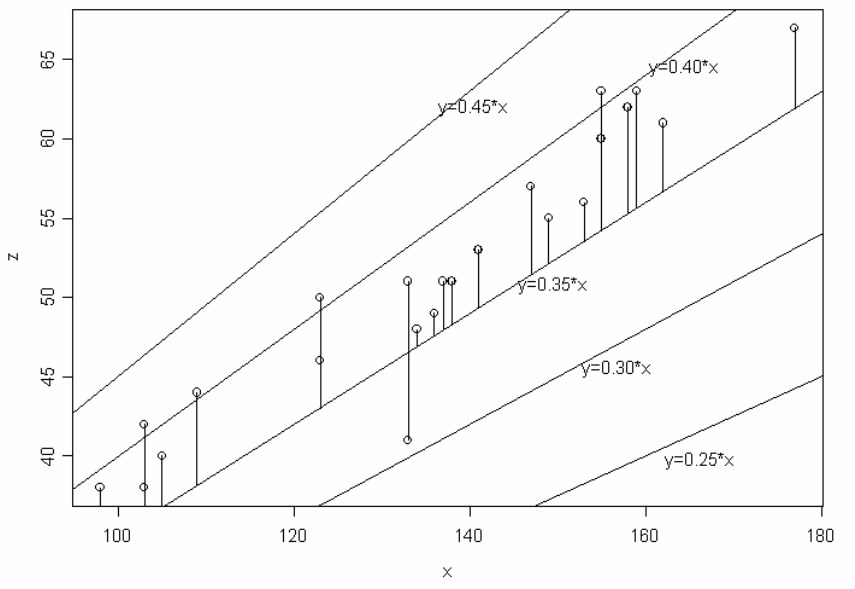
Théorème de Pythagore : La variance plus le carré de la moyenne égale la moyenne des carrés. On dit que la variance vaut la moyenne des carrés moins le carré de la moyenne.

2.2. Régression par l'origine

Trouver *a* qui minimise

$$E(a) = \sum_{i=1}^{n} p_i \left(y_i - a x_i\right)^2$$

La tortue croit en longueur et en hauteur. Change t-elle de forme ?

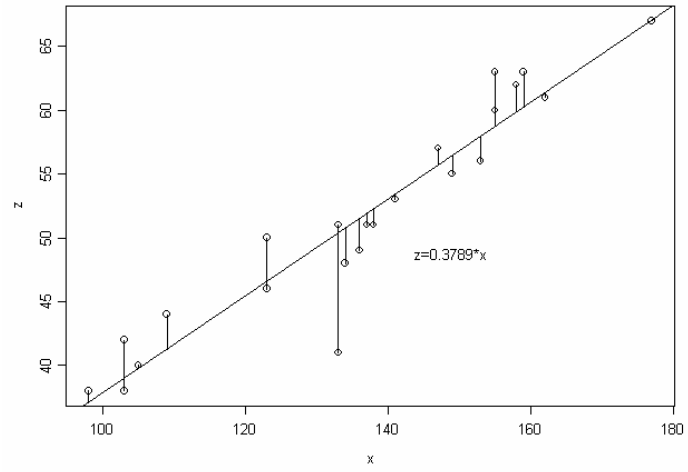



```
data(tortues)
x <- tortues$long
z <- tortues$haut
plot(x,z) ; abline(0,0.25)
text(locator(1),"y=0.25*x")
abline(0,0.30) ; text(locator(1),"y=0.30*x")
abline(0,0.35) ; text(locator(1),"y=0.35*x")
abline(0,0.40) ; text(locator(1),"y=0.40*x")
abline(0,0.45) ; text(locator(1),"y=0.45*x")
for (i in 1:25) segments(x[i],z[i],x[i],0.35   *x[i])
```

Application du théorème du pied de la perpendiculaire :

$$a = \frac{\langle \mathbf{x} | \mathbf{y} \rangle}{\langle \mathbf{x} | \mathbf{x} \rangle} = \frac{\sum_{i=1}^{n} p_i x_i y_i}{\sum_{i=1}^{n} p_i x_i^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$a = \frac{172142}{454286} = 0.3789$$

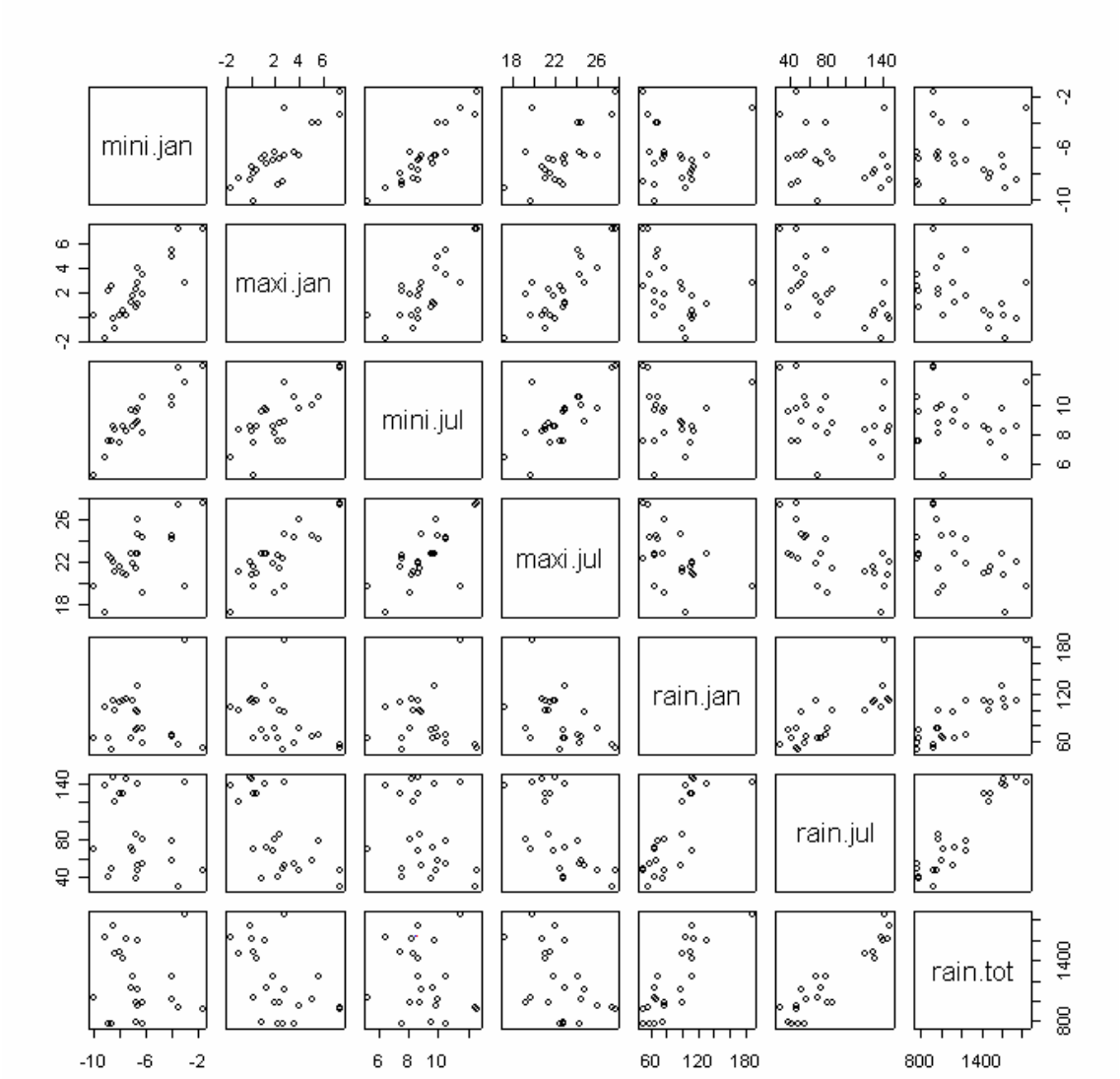


On ne peut pas faire mieux.

Fondamentalement *n* points de $\mathbb{R}^2$ sont représentés sur cette figure et l'ajustement se fait sur 2 points de $\mathbb{R}^n$.

2.3. Covariance et corrélation

```
data(atlas) ; plot(atlas$meteo)
```



$$m(\mathbf{x}) = \sum_{i=1}^{n} p_i x_i \text{ et } m(\mathbf{y}) = \sum_{i=1}^{n} p_i y_i$$

$$v(\mathrm{x}) = \sum_{i=1}^{n} p_i \left(x_i - m(\mathbf{x})\right)^2 \text{ et } v(\mathrm{y}) = \sum_{i=1}^{n} p_i \left(y_i - m(\mathbf{y})\right)^2$$

$$c(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} p_i \left(x_i - m(\mathbf{x})\right)\left(y_i - m(\mathbf{y})\right)$$

La covariance est un produit scalaire entre variables centrées :

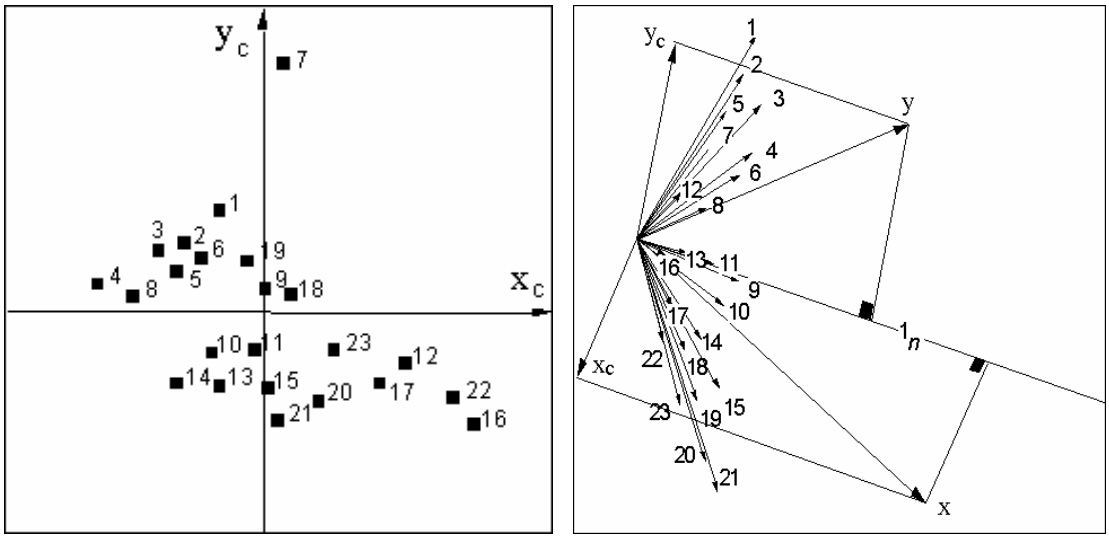
$$c(\mathbf{x},\mathbf{y}) = \left\langle \mathbf{x} - m(\mathbf{x})\mathbf{1}_n \middle| \mathbf{y} - m(\mathbf{y})\mathbf{1}_n \right\rangle = \left\langle \mathbf{x}_0 \middle| \mathbf{y}_0 \right\rangle$$

La corrélation est le cosinus de l'angle des deux variables centrées :

$$r(\mathbf{x},\mathbf{y}) = \frac{c(\mathbf{x},\mathbf{y})}{\sqrt{v(\mathbf{x})}\sqrt{v(\mathbf{y})}} = \frac{\left\langle \mathbf{x}_0 \middle| \mathbf{y}_0 \right\rangle}{\|\mathbf{x}_0\|\|\mathbf{y}_0\|} = \cos\left(\mathbf{x}_0, \mathbf{y}_0\right)$$

```
> round(cor(atlas$meteo),digits=2)
         mini.jan maxi.jan mini.jul maxi.jul rain.jan rain.jul rain.tot
mini.jan     1.00     0.83     0.91     0.60    -0.01    -0.25    -0.09
maxi.jan     0.83     1.00     0.78     0.82    -0.43    -0.64    -0.50
mini.jul     0.91     0.78     1.00     0.72    -0.03    -0.27    -0.12
maxi.jul     0.60     0.82     0.72     1.00    -0.50    -0.64    -0.50
rain.jan    -0.01    -0.43    -0.03    -0.50     1.00     0.78     0.84
rain.jul    -0.25    -0.64    -0.27    -0.64     0.78     1.00     0.93
rain.tot    -0.09    -0.50    -0.12    -0.50     0.84     0.93     1.00
> round(360*acos(cor(atlas$meteo))/2/pi,digits=0)
         mini.jan maxi.jan mini.jul maxi.jul rain.jan rain.jul rain.tot
mini.jan        0       34       24       53       90      105       95
maxi.jan       34        0       39       35      116      130      120
mini.jul       24       39        0       44       92      106       97
maxi.jul       53       35       44        0      120      130      120
rain.jan       90      116       92      120        0       39       33
rain.jul      105      130      106      130       39        0       21
rain.tot       95      120       97      120       33       21        0
```

La corrélation de min.jan et max.jan vaut 0.83. Les deux variables centrées font un angle de 34 degrés. La corrélation de max.ju let rain.jul vaut –0.64. Les deux variables centrées font un angle de 130 degrés.



A gauche, 2 variables centrées vues comme n points de $\mathbb{R}^2$. La covariance est négative. A droite, les mêmes variables vues comme 2 points de $\mathbb{R}^n$. L'angle est obtus (cosinus négatif).

Remarque :

$$\begin{aligned} c(\mathbf{x},\mathbf{y}) &= \left\langle \mathbf{x} - m(\mathbf{x})\mathbf{1}_n \middle| \mathbf{y} - m(\mathbf{y})\mathbf{1}_n \right\rangle \\ &= \left\langle \mathbf{x} \middle| \mathbf{y} \right\rangle - m(\mathbf{x})\left\langle \mathbf{1}_n \middle| \mathbf{y} \right\rangle - m(\mathbf{y})\left\langle \mathbf{x} \middle| \mathbf{1}_n \right\rangle + m(\mathbf{x})m(\mathbf{y})\left\langle \mathbf{1}_n \middle| \mathbf{1}_n \right\rangle \\ &= \left\langle \mathbf{x} \middle| \mathbf{y} \right\rangle - m(\mathbf{x})m(\mathbf{y}) \end{aligned}$$

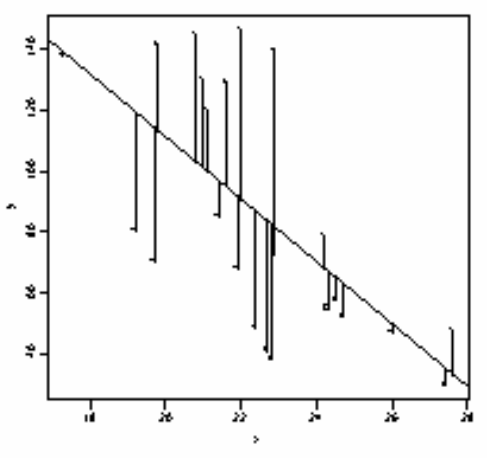
On dit que la covariance vaut la moyenne des produits moins le produit des moyennes.

2.4. Le problème de la régression simple

Trouver *a* et *b* pour minimiser la moyenne des carrées des écarts (critère des **moindres carrés**), soit minimiser :

$$E(a,b) = \sum_{i=1}^{n} p_i \left(y_i - (ax_i + b) \right)^2$$

```
> x <- atlas$meteo$maxi.jul
> y <- atlas$meteo$rain.jul
> plot(x,y)
> abline(lm(y~x))
> segments(x,predict(lm(y~x)),x,y)
```



2.5. Le théorème des moindres carrés

Sous-espace vectoriel
Si F est une partie de $E = \mathbb{R}^n$ telle que :

$$\begin{cases} \mathbf{x} \in F \text{ et } \mathbf{y} \in F \Rightarrow \mathbf{x} + \mathbf{y} \in F \\ \mathbf{x} \in F,\ \alpha \in \mathbb{R} \Rightarrow \alpha\mathbf{x} \in F \\ \mathbf{0} \in F \end{cases}$$

alors F est un espace vectoriel. On dit que F est un sous-espace vectoriel de E.

Combinaison linéaire
Si $\mathbf{v}_1$, ..., $\mathbf{v}_r$ sont *r* éléments de E, on peut considérer les vecteurs de E du type :

$$\mathbf{x} = \alpha_1\mathbf{v}_1 + \ldots + \alpha_r\mathbf{v}_r \quad \text{avec} \quad \alpha_1, \ldots, \alpha_r \in \mathbb{R}$$

On dit que $\mathbf{x}$ est une combinaison linéaire des vecteurs $\mathbf{v}_1$, ...,$\mathbf{v}_r$. Quand les α_k varient, $\mathbf{x}$ varie et on note :

$$H=\text{sev}(\mathbf{v}_1, \ldots, \mathbf{v}_r) = \{\mathbf{x} \in E / \mathbf{x} = \alpha_1\mathbf{v}_1 + \ldots + \alpha_r\mathbf{v}_r\}$$

C'est un sous-espace vectoriel de E. On dit que c'est le sous-espace vectoriel *engendré* par les vecteurs $\mathbf{v}_1$, ..., $\mathbf{v}_r$.

Indépendance linéaire
Si $\alpha_1\mathbf{v}_1 + \ldots + \alpha_r\mathbf{v}_r = \mathbf{0} \Rightarrow \alpha_1 = \ldots = \alpha_r = 0 \in \mathbb{R}$ les vecteurs $\mathbf{v}_1$, ...,$\mathbf{v}_r$ sont dits linéairement indépendants. On dit qu'ils forment un système libre. Dans le cas contraire, ils sont linéairement dépendants ou forment un système lié. Dans ce cas :

$$\exists \alpha_1, \ldots, \alpha_r \text{ non tous nuls tels que } \alpha_1\mathbf{v}_1 + \ldots + \alpha_r\mathbf{v}_r = \mathbf{0}$$

Générateur
Des vecteurs $\mathbf{v}_1$, ..., $\mathbf{v}_r$ forment un générateur de F, sous-espace vectoriel, si tout vecteur $\mathbf{v}$ de F s'écrit $\mathbf{v} = \alpha_1\mathbf{v}_1 + \ldots + \alpha_r\mathbf{v}_r$ avec $,\ldots,\alpha_r \in \mathbb{R}$. Ceci s'écrit encore :

$$\text{F=sev}\left(\mathbf{v}_1,\ldots,\mathbf{v}_r\right).$$

Base d'un sous-espace vectoriel
Des vecteurs $\mathbf{v}_1$, ..., $\mathbf{v}_r$ forment une base de F, sous-espace vectoriel, si ils forment un générateur libre. Tout sous-espace a au moins une base et si une base a r vecteurs, toutes les autres bases ont aussi r vecteurs. r est la dimension de F. Si on ajoute un vecteur à une base, le système obtenu est lié. Si on enlève un vecteur à une base le système obtenu n'est plus un générateur. Si $\mathbf{v}_1$, ..., $\mathbf{v}_r$ forment une base de F, l'écriture :

$$\mathbf{x} = \alpha_1\mathbf{v}_1 + \ldots + \alpha_r\mathbf{v}_r \quad \text{avec} \quad \alpha_1,\ldots,\alpha_r \in \mathbb{R}$$

est unique.

Base orthogonale
Si $\mathbf{v}_1$, ..., $\mathbf{v}_r$ est une base de F, on dit qu'elle est Φ–orthogonale si :

$$i \neq j \Rightarrow \left\langle \mathbf{v}_i \middle| \mathbf{v}_j \right\rangle_\Phi = 0$$

On dit qu'elle est Φ–normale si en plus $\left\|\mathbf{v}_i\right\|_\Phi = 1$ pour $1 \leq i \leq r$
Tout système $\mathbf{v}_1$, ..., $\mathbf{v}_r$ de r vecteurs non nuls et orthogonaux deux à deux est une base orthogonale du sous-espace qu'ils engendrent.

Orthogonalisation de Gram-Schmidt
Quand on possède une base $\mathbf{v}_1$, ..., $\mathbf{v}_r$ de F on peut toujours trouver une base orthogonale par le procédé suivant :

$$\text{pas } 1 \ \mapsto \ \mathbf{w}_1 = \mathbf{v}_1$$

$$\text{pas } 2 \ \mapsto \ \mathbf{w}_2 = \mathbf{v}_2 - \frac{\left\langle \mathbf{v}_2 \middle| \mathbf{w}_1 \right\rangle}{\left\langle \mathbf{w}_1 \middle| \mathbf{w}_1 \right\rangle}\mathbf{w}_1$$

$$\text{pas } 3 \ \mapsto \ \mathbf{w}_3 = \mathbf{v}_3 - \frac{\left\langle \mathbf{v}_3 \middle| \mathbf{w}_1 \right\rangle}{\left\langle \mathbf{w}_1 \middle| \mathbf{w}_1 \right\rangle}\mathbf{w}_1 - \frac{\left\langle \mathbf{v}_3 \middle| \mathbf{w}_2 \right\rangle}{\left\langle \mathbf{w}_2 \middle| \mathbf{w}_2 \right\rangle}\mathbf{w}_2$$

...

$$\text{pas } r \ \mapsto \ \mathbf{w}_r = \mathbf{v}_r - \frac{\left\langle \mathbf{v}_r \middle| \mathbf{w}_1 \right\rangle}{\left\langle \mathbf{w}_1 \middle| \mathbf{w}_1 \right\rangle}\mathbf{w}_1 - \ldots - \frac{\left\langle \mathbf{v}_r \middle| \mathbf{w}_{r-1} \right\rangle}{\left\langle \mathbf{w}_{r-1} \middle| \mathbf{w}_{r-1} \right\rangle}\mathbf{w}_{r-1}$$

Décomposition orthogonale
On considère une sous-espace F et un vecteur $\mathbf{y}$ non nul. Soit $\mathbf{w}_1$, ..., $\mathbf{w}_r$ une base orthogonale quelconque de F. On peut projeter $\mathbf{y}$ sur un des vecteurs $\mathbf{w}_k$:

$$P_{/\mathbf{w}_k}\left(\mathbf{y}\right) = \mathbf{y}_k = \frac{\left\langle \mathbf{y} \middle| \mathbf{w}_k \right\rangle}{\left\langle \mathbf{w}_k \middle| \mathbf{w}_k \right\rangle}\mathbf{w}_k$$

On peut sommer les vecteurs projetés, le résultat est un vecteur de F. On l'appelle :

$$\hat{\mathbf{y}}_{\text{F}} = \mathbf{y}_1 + \mathbf{y}_2 + \ldots + \mathbf{y}_r$$

Alors $\mathbf{y} = \hat{\mathbf{y}}_F + (\mathbf{y} - \hat{\mathbf{y}}_F)$ et $\mathbf{y} - \hat{\mathbf{y}}_F$ est orthogonal à chaque vecteur $\mathbf{w}_k$, donc à chacun des vecteur de F (on dit qu'il est orthogonal à F). On a décomposé **y** en un vecteur $\hat{\mathbf{y}}_F$ de F et un vecteur $\mathbf{y} - \hat{\mathbf{y}}_F$ orthogonal à F. Par définition, $\hat{\mathbf{y}}_F$ est dit le projeté de **y** sur F et est noté :

$$\hat{\mathbf{y}} = P_{/F}(\mathbf{y})$$

Le projeté de **y** sur F est **l'unique** vecteur **z** de F tel que $\mathbf{y} - \mathbf{z}$ est orthogonal à F. Si il y a deux **z** et $\hat{\mathbf{y}}_F$ alors :

$$\begin{array}{l} \mathbf{y} = \hat{\mathbf{y}}_F + (\mathbf{y} - \hat{\mathbf{y}}_F) \\ \mathbf{y} = \mathbf{z} + (\mathbf{y} - \mathbf{z}) \\ \hline \mathbf{0} = (\hat{\mathbf{y}}_F - \mathbf{z}) + (\mathbf{z} - \hat{\mathbf{y}}_F) \end{array}$$

$\hat{\mathbf{y}}_F - \mathbf{z}$ est orthogonal à lui-même donc nul et les deux vecteurs sont égaux.

Fondamentalement le vecteur projeté $\hat{\mathbf{y}}_F$ est le vecteur de F le plus proche de **y** :

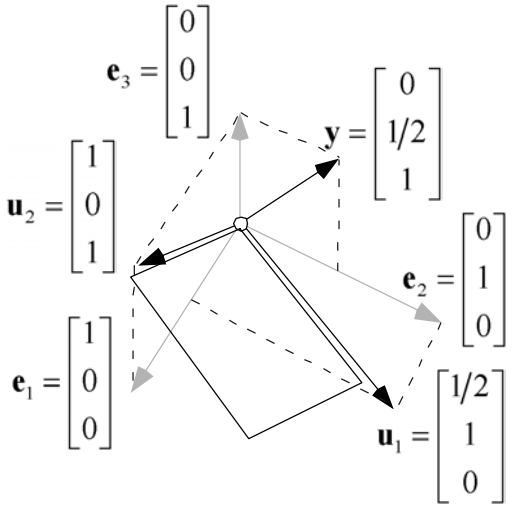
$$\mathbf{z} \in F \Rightarrow \|\mathbf{y} - \hat{\mathbf{y}}_F\|^2 \leq \|\mathbf{y} - \mathbf{z}\|^2$$

Enfin, la projection est une application linéaire :

$$\widehat{\mathbf{y}_1 + \mathbf{y}_2} = P_{/F}(\mathbf{y}_1 + \mathbf{y}_2) = P_{/F}(\mathbf{y}_1) + P_{/F}(\mathbf{y}_2) = \widehat{\mathbf{y}_1} + \widehat{\mathbf{y}_2}$$

$$\widehat{\alpha\mathbf{y}} = P_{/F}(\alpha\mathbf{y}) = \alpha P_{/F}(\mathbf{y}) = \alpha\hat{\mathbf{y}}$$

Exemple. Quelles sont les coordonnées de la projection de **y** sur le plan engendré par $\mathbf{u}_1$ et $\mathbf{u}_2$?



2.6. La solution de la régression simple

Chercher *a* et *b* qui minimise $E(a,b) = \sum_{i=1}^{n} p_i (y_i - (ax_i + b))^2$, c'est chercher *a* et *b* qui minimise $\|\mathbf{y} - a\mathbf{x} - b\mathbf{1}_n\|^2$, c'est chercher le projeté de **y** sur le plan π défini par **x** et $\mathbf{1}_n$.

Ce plan admet pour base orthogonale $\mathbf{x}_0 = \mathbf{x} - m(\mathbf{x})\mathbf{1}_n$ et $\mathbf{1}_n$. Le projeté est :

$$\hat{\mathbf{y}} = P_{/\pi}(\mathbf{y}) = P_{/\pi}(\mathbf{y}_0 + m(\mathbf{y})\mathbf{1}_n) = P_{/\pi}(\mathbf{y}_0) + P_{/\pi}(m(\mathbf{y})\mathbf{1}_n)$$

$$\hat{\mathbf{y}} = P_{/\mathbf{x}_0}(\mathbf{y}_0) + P_{/\mathbf{x}_0}(m(\mathbf{y})\mathbf{1}_n) + P_{/\mathbf{1}_n}(\mathbf{y}_0) + P_{/\mathbf{1}_n}(m(\mathbf{y})\mathbf{1}_n)$$

$$\hat{\mathbf{y}} = \frac{\langle \mathbf{y}_0 | \mathbf{x}_0 \rangle}{\langle \mathbf{x}_0 | \mathbf{x}_0 \rangle}\mathbf{x}_0 + 0 + 0 + m(\mathbf{y})\mathbf{1}_n$$

$$a = \frac{c(\mathbf{x},\mathbf{y})}{v(\mathbf{x})} \quad b = m(\mathbf{y}) - am(\mathbf{x})$$

Avec de plus :

$$m(\hat{\mathbf{y}}) = m(\mathbf{y})$$

$$\hat{\mathbf{y}}_0 = P_{/\mathbf{x}_0}(\mathbf{y}_0) = a\mathbf{x}_0$$

2.7. Décomposition de la variance

La moyenne des prédictions est égale à la moyenne des observations. Ce qui vient d'être vérifié dans la régression simple est très général.

Théorème des trois perpendiculaires

Si F est un sous-espace de E et G un sous-espace de F, alors :

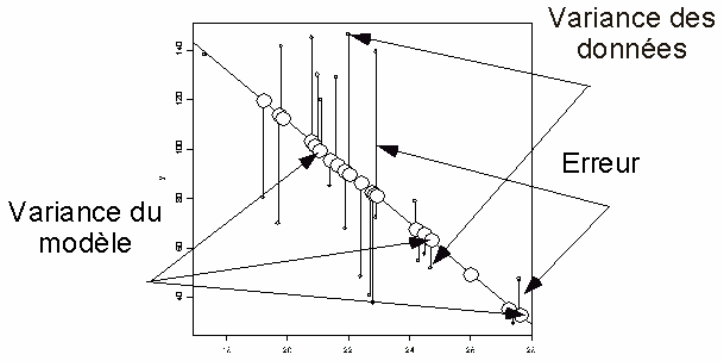
$$P_{/\mathrm{G}}(P_{/\mathrm{F}}(\mathbf{y})) = P_{/\mathrm{G}}(\mathbf{y})$$

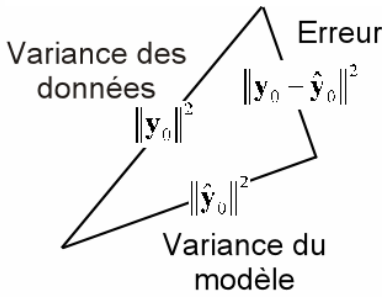
Donc, si le sous-espace de projection contient $\mathbf{1}_n$, alors :

$$m(\mathbf{y})\mathbf{1}_n = P_{/\mathbf{1}_n}(\mathbf{y}) = P_{/\mathbf{1}_n}(P_{/\pi}(\mathbf{y})) = P_{/\mathbf{1}_n}(\hat{\mathbf{y}}) = m(\hat{\mathbf{y}})\mathbf{1}_n$$

$$\left.\begin{aligned} \hat{\mathbf{y}}_0 &= P_{/\mathbf{x}_0}(\mathbf{y}_0) = a\mathbf{x}_0 \\ \mathbf{y}_0 &= \hat{\mathbf{y}}_0 + (\mathbf{y}_0 - \hat{\mathbf{y}}_0) \end{aligned}\right] \Rightarrow \mathrm{var}(\mathbf{y}) = \mathrm{var}(\hat{\mathbf{y}}) + E(a,b)$$

A retenir les deux points de vue sur ces opérations :





3. Équation d'analyse de la variance

Le principe général s'étend aux variables qualitatives en introduisant les indicatrices des classes.

A	B	C	D			E				F
22.9	0.15	1	1000							-0.695
20.8	-0.7	1	1000		1	2	3	4		-0.695
22	-0.215	1	1000		0.15	-0.457	0.676	0.15		-0.695
17.3	-2.116	1	1000		-0.7	0.109	2.051	0.797		-0.695
21.6	-0.376	1	1000		-0.215	-1.348	1.97	0.878		-0.695
21	-0.619	1	1000		-2.116	-1.145		-0.255		-0.695
19.8	-1.105	1	1000		-0.376	0.069		1.404		-0.695
21.1	-0.579	1	1000		-0.619	0.716				-0.695
21.4	-0.457	2	0100		-1.105	-0.053				-0.301
22.8	0.109	2	0100		-0.579					-0.301
19.2	-1.348	2	0100							-0.301
24.2	0.676	3	0010		8	7	3	5		1.566
22.9	0.15	4	0001		-0.695	-0.301	1.566	0.595		0.595
19.7	-1.145	2	0100							-0.301
22.7	0.069	2	0100							-0.301
27.6	2.051	3	0010							1.566
24.5	0.797	4	0001							0.595
24.7	0.878	4	0001							0.595
21.9	-0.255	4	0001							0.595
24.3	0.716	2	0100							-0.301
22.4	-0.053	2	0100							-0.301
27.4	1.97	3	0010							1.566
26	1.404	4	0001							0.595

A Une variable observée (données brutes) . B la variable normalisée (données transformées de moyenne 0 et de variance 1). C Une variable qualitative enregistrée par numéros de modalités. D La variable qualitative vue comme l'ensemble de ses indicatrices de classe. E Tableau d'analyse de variance (les valeurs sont rangées, comptées et moyennées par classe). F Le modèle de la variable normalisée (une valeur est modélisée par la moyenne de sa classe).

Une variable **qualitative** est un enregistrement prenant ses valeurs dans un ensemble de possibilités appelées **modalités**. Le code des modalités (du type 1 = « bleu », 2 = « vert », …) donne un numéro d'ordre à chaque modalité qui permet l'enregistrement sous forme d'entier. Quand l'ordre des modalités a un sens (du type « un peu », « beaucoup », « passionnément », …) on dit que la variable est **qualitative à modalité ordonnée** ou **semi-quantitative**. On note **q** cette variable. Elle prend les valeurs q_i et on note indifféremment $q_i = k$ (**q** prend la valeur *k* pour l'individu *i*) ou $i \in \mathbf{q}_k$ (l'individu *i* appartient à la classe *k* de la variable **q**) ou $Cl_{\mathbf{q}}(i) = k$ (la classe de l'individu *i* est *k*).

Si *n* est le nombre total de mesures, n_k est le nombre d'individus porteurs de la modalité *k* (*k* variant de 1 à *m*, nombre total de modalités). Si les individus portent les poids p_i, le poids d'une classe est $p_k^+ = \sum_{q_i=k} p_i$. Pour la pondération uniforme (le cas le plus courant, suffisant pour la suite) on a $p_k^+ = \frac{n_k}{n}$.

On note $\mathbf{I}_k$ le vecteur à *n* composantes égales à 1 si $q_i = k$ et 0 sinon. Ce vecteur est l'**indicatrice** de la classe *k*.

L'ensemble des vecteurs $\mathbf{I}_1,...,\mathbf{I}_k,...,\mathbf{I}_m$ et la variable **q** contiennent donc exactement la même information. Au plan théorique, c'est la présentation par les indicatrices qui joue un rôle fondamental. On utilise la métrique uniforme dans $\mathbb{R}^n$ ($\langle \mathbf{x}|\mathbf{y} \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$).

Le carré de la norme d'une indicatrice est son poids :

$$\|\mathbf{I}_k\|^2 = \frac{1}{n}\left(\sum_{\substack{i=1 \\ q_i \neq k}}^{n} 0^2 + \sum_{\substack{i=1 \\ q_i = k}}^{n} 1^2 \right) = \frac{n_k}{n} = p_k^+$$

Le produit scalaire de deux indicatrices est nul :

$$\langle \mathbf{I}_k | \mathbf{I}_j \rangle = \frac{1}{n}\left(\sum_{i=1}^{n} \partial_{ik}\partial_{jk} \right) = 0$$

puisque la somme ne contient que des couples 00 ou 01 ou 10. Les indicatrices sont **orthogonales**.

La somme des indicatrices vaut $\sum_{k=1}^{m} \mathbf{I}_k = \mathbf{1}_n$.

Considérons alors une variable quantitative **x** .

$$\langle \mathbf{x}|\mathbf{I}_k \rangle = \frac{1}{n}\left(\sum_{\substack{i=1 \\ q_i \neq k}}^{n} 0x_i + \sum_{\substack{i=1 \\ q_i = k}}^{n} 1x_i \right) = \frac{n_k}{n}\frac{1}{n_k}\sum_{\substack{i=1 \\ q_i = k}}^{n} x_i = p_k^+ m(\mathbf{x}/\mathbf{q}=k)$$

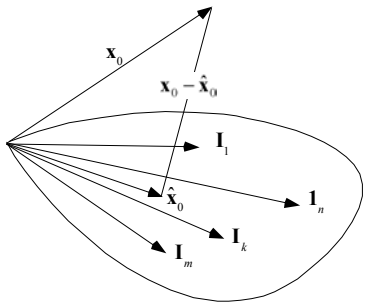
où $m(\mathbf{x}/\mathbf{q}=k)$ est la moyenne des valeurs de **x** dans la classe *k* (on dit moyenne conditionnelle de **x** sachant que **q** vaut *k*). Donc :

$$P_{/\mathbf{I}_k}(\mathbf{x}) = \frac{\langle \mathbf{x}|\mathbf{I}_k \rangle}{\langle \mathbf{I}_k|\mathbf{I}_k \rangle}\mathbf{I}_k = m(\mathbf{x}/\mathbf{q}=k)\mathbf{I}_k$$

Donc si H est le sous-espace engendré par l'ensemble des indicatrices, donc l'ensemble des combinaisons linéaires d'indicatrices ou encore l'ensemble des variables constantes par classe :

$$P_{/H}(\mathbf{x}) = \sum_{k=1}^{m} m(\mathbf{x}/\mathbf{q}=k)\mathbf{I}_k$$

La variable constante par classe la plus proche de **x** est obtenue en remplaçant chaque valeur par la moyenne de la classe correspondante. C'est vrai aussi pour la variable centrée $\mathbf{x}_0 = \mathbf{x} - m(\mathbf{x})\mathbf{1}_n$:



On sait que le carré de la longueur de la variable centrée est la variance de **x**. La variable projetée est aussi centrée et sa variance vaut :

$$v(\hat{\mathbf{x}}_0) = \frac{1}{n}\sum_{i=1}^{n}\left(m(\mathbf{x}_0 / \mathbf{q} = Cl(i))\right)^2 = \sum_{k=1}^{m}\frac{n_k}{n}\left(m(\mathbf{x}_0 / \mathbf{q} = k)\right)^2 = \sum_{k=1}^{m} p_k^+ \left(m(\mathbf{x}_0 / \mathbf{q} = k)\right)^2$$

C'est la moyenne des carrés des moyennes par classe de la variable centrée, donc la variance des moyennes par classe, ce qu'on appelle la **variance inter-classe**.

L'écart entre les deux a aussi un sens :

$$\begin{aligned}\|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - m(\mathbf{x} / \mathbf{q} = Cl(i))\right)^2 \\ &= \frac{1}{n}\sum_{k=1}^{m}\sum_{\substack{i=1 \\ \mathbf{q}_i = k}}^{n}\left(x_i - m(\mathbf{x} / \mathbf{q} = k)\right)^2 = \sum_{k=1}^{m} p_k^+ \left(v(\mathbf{x} / \mathbf{q} = k)\right)^2\end{aligned}$$

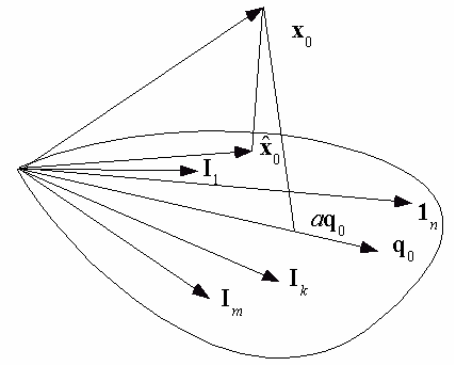
C'est la moyenne des variances conditionnelles (variance des valeurs qui sont dans la même classe), qui s'appelle **variance intra-classe.**

Variance totale = Variance Inter + Variance Intra est l'équation d'analyse de la variance, autre cas particulier du théorème de Pythagore $\|\mathbf{x}_0\|^2 = \|\hat{\mathbf{x}}_0\|^2 + \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|^2$.

Le rapport de corrélation (rapport de la variance inter sur la variance totale) n'est rien d'autre que le carré du cosinus de l'angle entre la variable centrée et le sous-espace :

$$\eta^2(\mathbf{x}, \mathbf{q}) = \frac{\text{Variance Inter}}{\text{Variance Totale}} = \frac{\|\hat{\mathbf{x}}_0\|^2}{\|\mathbf{x}_0\|^2} = \cos^2 A(\mathbf{x}_0, \hat{\mathbf{x}}_0) = \cos^2 A(\mathbf{x}_0, H)$$

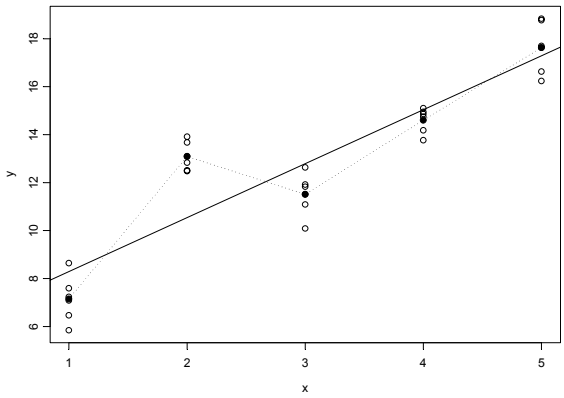
Remarque 1 : quand **q est semi-quantitative** elle définit l'ensemble des indicatrices d'une part mais peut être aussi considérée comme une variable quantitative constante par classe située donc dans le sous-espace des indicatrices :



La géométrie des variables indique alors qu'on a toujours :

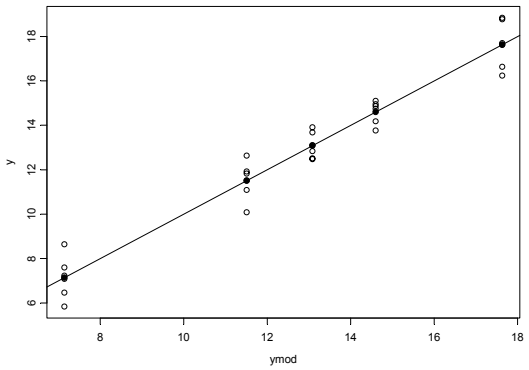
$$\eta^2(\mathbf{x}, \mathbf{q}) \geq r^2(\mathbf{x}, \mathbf{q})$$

En effet, le vecteur du sous-espace le plus proche de la variable centrée est certainement plus proche que le vecteur le plus proche porté par l'explicative (exactement comme le champion du monde est meilleur que le champion du quartier, voire, mais c'est rare, égal). Ceci signifie que le modèle par la droite de régression est toujours moins bon que le modèle par la courbe régression :



*95 % de la variance de **y** est expliquée par la courbe de régression, 80% par la droite de régression.*

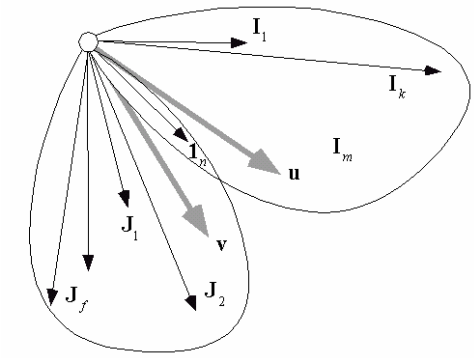
Remarque 2 : Pour faire correspondre la droite et la courbe, il suffit de positionner une modalité à la moyenne des valeurs de la classe. La droite de régression est la courbe de régression. Le rapport de corrélation devient le carré du coefficient de corrélation.



Ceci est une pratique de **scoring** (association d'un score numérique aux modalités d'une variable qualitative dans un but précis).

Remarque 3 : Ce paragraphe sur le rapport de corrélation est essentiel pour comprendre l'**analyse des correspondances multiples** équivalent pour les variables qualitatives de l'ACP.

Remarque 4 :



Quand on a deux variables qualitatives donc deux espaces engendrés par deux paquets d'indicatrices, le vecteur $\mathbf{1}_n$ est commun aux deux espaces. On peut chercher (et trouver) le vecteur **u** orthogonal à $\mathbf{1}_n$ dans le premier et le vecteur **v** orthogonal à $\mathbf{1}_n$ dans le second qui minimisent l'angle qu'ils font entre eux. C'est un problème d'**analyse canonique** en général et la base de l'**analyse des correspondances** dans ce cas particulier.

4. Régression multiple

4.1. Position du problème : régression prédictive

L'objectif est de prédire une variable **y** avec plusieurs variables qui forment un tableau **X**. La colonne *j* de **X** est notée $\mathbf{x}^j$. **X** contient des variables **explicatives**. Les moyennes et variances des explicatives sont :

$$m_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \left\langle \mathbf{x}^j \middle| \mathbf{1}_n \right\rangle$$

$$v_j = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - m_j\right)^2 = \left\|\mathbf{x}^j - m_j\mathbf{1}_n\right\|^2 = \left\|\mathbf{x}_0^j\right\|^2$$

L'objectif est de trouver le meilleur modèle du type :

$$\hat{y}_i = a_1 x_{i1} + \cdots + a_p x_{ip} + b$$

On cherche à minimiser :

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \left(a_1 x_{i1} + \cdots + a_p x_{ip} + b\right)\right)^2$$

c'est chercher un vecteur du type :

$$\hat{\mathbf{y}} = a_1\mathbf{x}^1 + \cdots + a_p\mathbf{x}^p + b\mathbf{1}_n$$

qui minimise :

$$\left\|\mathbf{y} - \hat{\mathbf{y}}\right\|^2$$

dans le sous-espace vectoriel engendré par $\mathbf{1}_n$ et les colonnes de **X**, c'est-à-dire le sous-espace $\mathrm{F} = \mathrm{sev}\left(\mathbf{x}^1, \ldots \mathbf{x}^p, \mathbf{1}_n\right)$.

Variables sans redondance :

$$\mathrm{F} = \mathrm{sev}\left(\mathbf{x}^1, \ldots, \mathbf{x}^p\right) + \Delta_n = \mathrm{sev}(\mathbf{X}) + \Delta_n$$

$$\mathrm{F} = \mathrm{sev}(\mathbf{X}) \oplus \Delta_n$$

Si le rang de **X** est égal à *p*, on dira que l'ensemble des variables explicatives *est sans redondance*.

Les vecteurs $\mathbf{x}^1,...,\mathbf{x}^p$ et $\mathbf{1}_n$ sont indépendants si et seulement si les vecteurs $\mathbf{x}_0^1,...,\mathbf{x}_0^p$ le sont.

Pour qu'un ensemble de variables explicatives soit sans redondance il faut et il suffit que la matrice des corrélations ou des covariances associée soit inversible.

$$\mathbf{C} = \left[c\left(\mathbf{x}^j, \mathbf{x}^k\right)\right] = \left[\left\langle \mathbf{x}_0^j \middle| \mathbf{x}_0^k \right\rangle \right] = \frac{1}{n}\left[\sum_{i=1}^{n} \left(x_{ij} - m_j\right)\left(x_{ik} - m_k\right) \right] = \frac{1}{n}\mathbf{X}_0^{\mathrm{t}}\mathbf{X}_0$$

$$\mathbf{R} = \left[r\left(\mathbf{x}^j, \mathbf{x}^k\right)\right] = \left[\frac{c\left(\mathbf{x}^j, \mathbf{x}^k\right)}{\sqrt{v_j}\sqrt{v_k}} \right] = \frac{1}{n}\mathbf{X}_*^{\mathrm{t}}\mathbf{X}_*$$

Si les variables explicatives sont sans redondance, le meilleur modèle linéaire au sens des moindres carrés :

$$\hat{\mathbf{y}} = a_1\mathbf{x}^1 + \cdots + a_p\mathbf{x}^p + b\mathbf{1}_n$$

s'écrit de manière unique. La solution est :

$$\mathbf{a} = \begin{bmatrix} a_1 \\ \cdots \\ a_j \\ \cdots \\ a_p \end{bmatrix} = \mathbf{C}^{-1} \begin{bmatrix} c\left(\mathbf{x}^1, \mathbf{y}\right) \\ \\ c\left(\mathbf{x}^j, \mathbf{y}\right) \\ \\ c\left(\mathbf{x}^p, \mathbf{y}\right) \end{bmatrix} \text{ et } b = m_y - \sum_{j=1}^{p} a_j m_j$$

La démonstration s'appuie sur la décomposition en quatre morceaux. D'une part, projeter sur H=sev$\left(\mathbf{x}^1,...,\mathbf{x}^p,\mathbf{1}_n\right)$ c'est projeter sur H=sev$\left(\mathbf{x}_0^1,...,\mathbf{x}_0^p,\mathbf{1}_n\right)$, c'est projeter sur H_0=sev$\left(\mathbf{x}_0^1,...,\mathbf{x}_0^p\right)$ et sur K=sev$\left(\mathbf{1}_n\right)$ puis ajouter les deux résultats (espaces orthogonaux). D'autre part, projeter **y** c'est projeter $\mathbf{y}_0$ et $m\left(\mathbf{y}\right)\mathbf{1}_n$ puis ajouter les deux résultats (la projection est linéaire). D'où :

$$\hat{\mathbf{y}} = P_{/\mathrm{H}}\left(\mathbf{y}_0\right) + P_{/\mathrm{K}}\left(\mathbf{y}_0\right) + P_{/\mathrm{H}}\left(m\left(\mathbf{y}\right)\mathbf{1}_n\right) + P_{/\mathrm{K}}\left(m\left(\mathbf{y}\right)\mathbf{1}_n\right)$$

Les deux termes centraux sont nuls par orthogonalité et on obtient le résultat avec le théorème général :

Si **X** contient en colonne *p* vecteurs de $\mathbb{R}^n$ indépendants et si **D** est un produit scalaire, la matrice du projecteur orthogonal sur le sous-espace engendré par les colonnes de **X** est $\mathbf{X}\left(\mathbf{X}^t\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{D}$.

Cas particulier : si **X** = **x**, c'est-à-dire si il n'y a qu'une variable explicative, alors :

$$b = m_y - a m_x \text{ et } a = \frac{c(x,y)}{c(x,x)} = \frac{c(x,y)}{v(x)}$$

4.2. Procédure

Structure d'un programme de régression prédictive multiple :

1) Calcul des moyennes des variables explicatives et centrage du tableau de ces variables : $\mathbf{X}_0 = \left[x_{ij} - m_j \right]$

2) Calcul de la matrice des covariances des variables explicatives :

$$\mathbf{C} = \frac{1}{n} \mathbf{X}_0^t \mathbf{X}_0$$

3) *Inversion* de **C**

4) Calcul de la moyenne de la variable à expliquer et centrage : $\mathbf{y}_0 = [y_i - m_y]$

5) Calcul du vecteur des covariances de la variable à expliquer et des variables explicatives : $\mathbf{d} = \frac{1}{n} \mathbf{X}_0^t \mathbf{y}_0$.

6) Calcul des coefficients de régression : $\mathbf{a} = \mathbf{C}^{-1}\mathbf{d}$

7) Calcul de l'ordonnée à l'origine : $b = m_y - \sum_{j=1}^{p} a_j m_j$

8) Calcul des valeurs prédites : $\hat{y}_i = a_1 x_{i1} + \cdots + a_p x_{ip} + b$

9) Calcul des résidus $r_i = y_i - \hat{y}_i$

10) Calcul des prédictions pour des valeurs supplémentaires :

$$\hat{y}_s = a_1 x'_{s1} + \cdots + a_p x'_{sp} + b$$

Cette procédure est souvent utilisée pour estimer des valeurs manquantes.

4.3. Carré de corrélation multiple

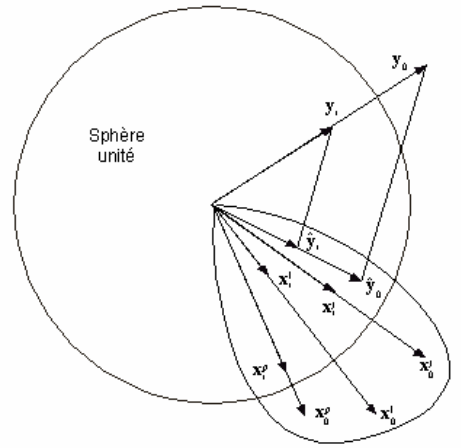
L'interprétation d'un modèle du type $\hat{y}_i = a_1 x_{i1} + \cdots + a_p x_{ip} + b$ est toujours un exercice difficile. En effet, on ne peut pas s'appuyer sur la valeur des coefficients qui dépendent des unités. Par exemple, un modèle du type :

abondance de l'espèce = 2 * distance à la source (km) – 3 * vitesse du courant (m/s)

s'écrit aussi bien :

abondance de l'espèce = 2000 * distance à la source (m) – 3 * vitesse du courant (m/s)

Ce qu'on aimerait savoir, c'est l'importance dans le résultat de chacune des explicatives (ce qui logiquement ne doit pas dépendre des unités de mesure !).



Pour décrire la qualité de la prédiction linéaire, on utilise la même stratégie que dans la régression simple. Le terme constant est ici inutile. En effet :

$$\hat{\mathbf{y}} = P_{/\mathrm{H}}(\mathbf{y}_0) + m(\mathbf{y})\mathbf{1}_n \Rightarrow m(\mathbf{y}) = m(\hat{\mathbf{y}}) \Rightarrow \hat{\mathbf{y}}_0 = P_{/\mathrm{H}}(\mathbf{y}_0) = \widehat{\mathbf{y}_0}$$

Simplement, la moyenne du modèle est la moyenne des données et le modèle de la variable centrée est le modèle centré. Pour éliminer les problèmes d'échelle, on normalise les données. On discute sur les variables :

$$\mathbf{y}_* = \frac{\mathbf{y}_0}{\|\mathbf{y}_0\|} = \left[\frac{y_i - m(\mathbf{y})}{\sqrt{v(\mathbf{y})}}\right] \text{ et } \mathbf{x}_*^k = \frac{\mathbf{x}_0^k}{\|\mathbf{x}_0^k\|} = \left[\frac{x_{ij} - m_j}{s_j}\right]$$

La normalisation ne change en rien le sous-espace de projection, la normalisation ne change que la longueur du vecteur projeté et $\widehat{\mathbf{y}_*} = \dfrac{\widehat{\mathbf{y}_0}}{\|\mathbf{y}_0\|} = \hat{\mathbf{y}}_*$ (le modèle normalisé est le normalisé du modèle). On retrouve facilement les valeurs des paramètres dans le modèle explicite en fonction des valeurs dans le modèle normalisé en comparant les écritures :

$$y_i = b + \sum_{k=1}^{p} a_k x_{ik} \text{ et } \frac{y_i - \overline{y}}{s_y} = \sum_{k=1}^{p} w_k \frac{x_{ik} - m_k}{s_k}$$

Le théorème de Pythagore donne :

$$\|\mathbf{y}_0\|^2 = v(\mathbf{y}) = v(\mathbf{y})\left\|\mathrm{P}_{/\mathrm{H}}(\mathbf{y}_0)\right\|^2 + E\left(a_1, \ldots, a_p, b\right) = v(\mathbf{y}) R^2_{y;x_1,\ldots,x_p} + E\left(a_1, \ldots, a_p, b\right)$$

La variance de **y** se décompose en la somme d'une erreur de prédiction et d'une partie de la variance dite variance expliquée. Donc $R^2_{y;x_1,\ldots,x_p}$ *est le pourcentage de variance expliquée*, ce qui étend le cas de la régression simple.

La quantité :

$$R^2_{y;x_1,\ldots,x_p} = \frac{\|\hat{\mathbf{y}}_0\|^2}{\|\mathbf{y}_0\|^2} = \frac{\text{Variance expliquée}}{\text{Variance totale}} = \frac{\|\hat{\mathbf{y}}_*\|^2}{1}$$

est appelée **carré de corrélation multiple** et généralise la mesure de corrélation simple. C'est au choix un pourcentage de variance expliquée ou le **carré du cosinus** de l'angle entre la variable à prédire et le sous-espace de prédiction.

Ceci explique deux phénomènes essentiels. Le premier est qu'**il faut éviter des explicatives corrélées** car le sous-espace de projection est instable numériquement. Ceci se comprend en posant une feuille de papier sur deux crayons. Si les deux crayons (deux explicatives) font un angle petit (corrélation grande) le moindre geste (une toute petite variation d'un vecteur) fait tomber la feuille (change complètement le plan de projection). Si les deux crayons font un angle droit (corrélation faible) le moindre geste (une toute petite variation d'un vecteur) n'a aucun effet (ne modifie pas sensiblement le plan de projection). Le second est qu'**il suffit d'ajouter des variables explicatives pour améliorer le modèle**. En effet une variable quelconque est plus près d'un plan que d'une droite qui contient ce plan, d'un sous-espace de dimension 3 que d'un sous-espace de dimension 2 qui est dedans …

$$\mathrm{A} \subseteq \mathrm{B} \Rightarrow \left\|P_{\mathrm{A}}(\mathbf{y})\right\|^2 \leq \left\|P_{\mathrm{B}}(\mathbf{y})\right\|^2$$

On peut donc explorer avec le carré de corrélation multiple le pouvoir de prédiction de chacune des variables, de chaque couple de variables, de chaque triplet de variables … On cherche alors un équilibre entre la **précision** du modèle et la **parcimonie** (l'économie du nombre de variables).

Illustration numérique (prédiction des 7 variables météorologiques par les trois variables géographiques – données de l'introduction).

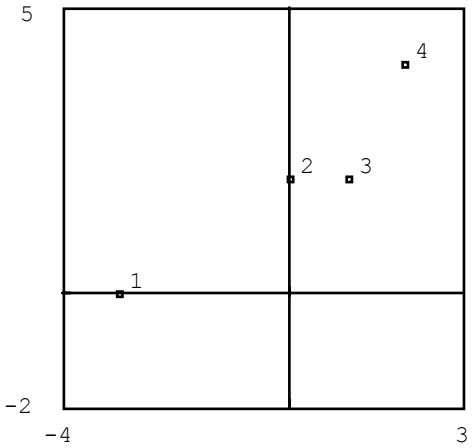
Variables explicatives			Variables à expliquer						
X	Y	Z	1	2	3	4	5	6	7
•	•	•	0.57	0.64	0.54	0.64	0.47	0.83	0.71
•	•		0.55	0.63	0.39	0.47	0.45	0.82	0.69
•		•	0.57	0.47	0.50	0.37	0.20	0.30	0.23
	•	•	0.52	0.64	0.53	0.60	0.47	0.83	0.70
•			0.55	0.47	0.39	0.33	0.03	0.09	0.01
	•		0.15	0.50	0.14	0.40	0.40	0.76	0.53
		•	0.47	0.28	0.49	0.35	0.01	0.00	0.05

La pluviométrie dépend d'abord de la latitude (Nord-Sud) mais la température dépend d'abord de l'altitude.

4.4. Exercice

La pondération implicite dans toute la suite est la pondération uniforme. Tous les résultats seront donnés sans approximation numérique. Tracer les solutions trouvées sur la figure. La variable **x** prend $n = 4$ valeurs : $\mathbf{x} = (-3, 0, 1, 2)$. La variable **y** prend $n = 4$ valeurs : $\mathbf{y} = (0, 2, 2, 4)$.

-> La moyenne de **x** vaut $m(\mathbf{x}) =$
-> Sa variance vaut $v(\mathbf{x}) =$
-> La moyenne de **y** vaut $m(\mathbf{y}) =$
-> Sa variance vaut $v(\mathbf{y}) =$
-> La covariance des deux variables vaut $c(\mathbf{x}, \mathbf{y}) =$
-> La corrélation des deux variables vaut $r(\mathbf{x}, \mathbf{y}) =$



Donner l'équation de la fonction $y = bx$ qui minimise l'erreur $E(b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - bx_i)^2$.

Donner l'équation de la fonction $y = bx + c$ qui minimise l'erreur $E(b,c) = \frac{1}{n}\sum_{i=1}^{n}(y_i - bx_i - c)^2$.

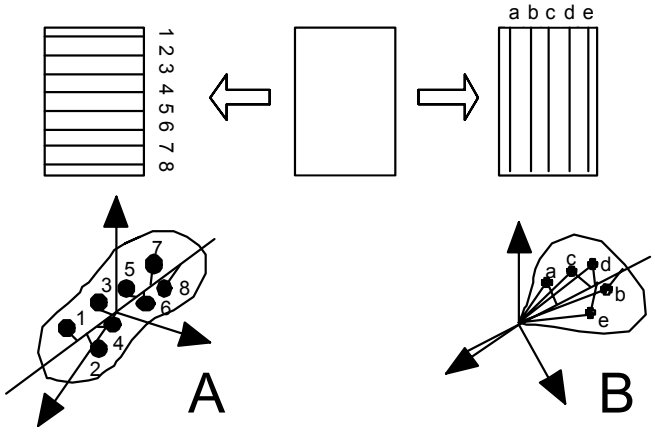
Donner l'équation de la fonction $x = by + c$ qui minimise l'erreur $E(b,c) = \frac{1}{n}\sum_{i=1}^{n}(x_i - by_i - c)^2$.

Donner l'équation de la fonction $y = ax^2 + bx + c$ qui minimise l'erreur $E(a,b,c) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - c\right)^2$.

Quel est le modèle qui donne l'erreur minimum ?

5. Composantes principales

Une dernière application de cette conception de la géométrie des variables concerne l'ACP déjà rencontrée. Historiquement, c'est le point de vue de Hotelling **1**. Nous avons vu la projection d'un nuage de points sur des axes qui maximisent l'inertie projetée. Le même problème a une autre signification dans l'espace des variables. Considérons un tableau de variables quantitatives **X** et le tableau normalisé $\mathbf{X}_*$. C'est aussi bien un ensemble de lignes (*n* points de $\mathbb{R}^p$) qu'un ensemble de colonnes (*p* points de $\mathbb{R}^n$).



Si **y** est une variable quelconque, on peut calculer sa corrélation avec chacune des variables de départ $r\left(\mathbf{y},\mathbf{x}^k\right)$ comme on peut calculer les corrélations entre variables initiales $r\left(\mathbf{x}^j,\mathbf{x}^k\right)$. Le lien entre **y** et **X** peut se mesurer par :

$$L\left(\mathbf{y},\mathbf{X}\right)=\sum_{k=1}^{p} r^2\left(\mathbf{y},\mathbf{x}^k\right)$$

Existe-t-il une variable **y** qui optimise cette quantité ? On peut supposer **y** de moyenne nulle et de variance 1 sans changer la question. $r^2\left(\mathbf{y},\mathbf{x}^k\right)$ est le carré de la norme du projeté de **y** sur le vecteur $\mathbf{x}_*^k$ aussi bien que le carré de la norme du projeté de $\mathbf{x}_*^k$ sur le vecteur **y**. Le lien est alors l'inertie projetée du nuage des variables sur **y** en pensant que le poids de chaque variable est 1 et que le produit scalaire de $\mathbb{R}^n$ est $\mathbf{D}=\left(1/n\right)\mathbf{I}_n$.

Nous avons résolu ce problème dans $\mathbb{R}^p$ en cherchant à maximiser l'inertie projetée du nuage de points où chaque point a le poids $1/n$ avec la métrique canonique.

La solution du nouveau problème suit le même principe. La seule différence est qu'ici nous travaillons sur un tableau normalisé : on dit qu'on pratique l'**analyse en composantes principales normée**.

1 Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* : 24, 417-441 , 498-520.

Rappelons les résultats (fiche Représentation des données multidimensionnelles). On cherche les axes principaux qui sont les vecteurs propres de $\mathbf{R} = \frac{1}{n}\mathbf{X}_*^t\mathbf{X}_*$ qui est exactement la matrice des corrélations des variables de **X**. Nous projetons ensuite les *n* points sur les axes principaux pour obtenir les coordonnées :

$$\mathbf{R} = \frac{1}{n}\mathbf{X}_*^t\mathbf{X}_* = \mathbf{U}\Lambda\mathbf{U}^t \text{ et } \mathbf{L} = \mathbf{X}\mathbf{U}$$

Maintenant l'inertie projetée sur un vecteur normé **y** est (1 pour montrer le poids) :

$$L(\mathbf{y},\mathbf{X}) = \sum_{k=1}^{p} r^2\left(\mathbf{y},\mathbf{x}_*^k\right) = \sum_{k=1}^{p} 1\left\langle \mathbf{y}\middle|\mathbf{x}_*^k\right\rangle^2 = \left(\frac{1}{n}\mathbf{X}_*^t\mathbf{y}\right)^t\left(\frac{1}{n}\mathbf{X}_*^t\mathbf{y}\right) = \frac{1}{n^2}\mathbf{y}^t\mathbf{X}_*\mathbf{X}_*^t\mathbf{y}$$

Le problème est très voisin. Une nuance : on cherche **y** normé pour le produit scalaire **D** soit sous la contrainte $\frac{1}{n}\sum_{i=1}^{n} y_i^2 = 1 \Leftrightarrow \sum_{i=1}^{n}\left(\frac{1}{\sqrt{n}}y_i\right)^2 = 1$. Posons donc $\mathbf{z} = \frac{1}{\sqrt{n}}\mathbf{y}$. Il faut trouver **z normé pour le produit scalaire ordinaire qui maximise** :

$$\frac{1}{n}\mathbf{z}^t\mathbf{X}_*\mathbf{X}_*^t\mathbf{z}$$

Comme le vecteur **u** normé pour le produit scalaire ordinaire qui maximise :

$$\frac{1}{n}\mathbf{z}^t\mathbf{X}_*^t\mathbf{X}_*\mathbf{z} = \mathbf{z}^t\mathbf{R}\mathbf{z}$$

est le premier vecteur propre de **R**, le vecteur **z** qui maximise $\frac{1}{n}\mathbf{z}^t\mathbf{X}_*\mathbf{X}_*^t\mathbf{z}$ est le premier vecteur propre de $\mathbf{S} = \frac{1}{n}\mathbf{X}_*\mathbf{X}_*^t$. Cette matrice a *n* lignes et *n* colonnes mais il n'est pas nécessaire de faire une nouvelle diagonalisation. En effet :

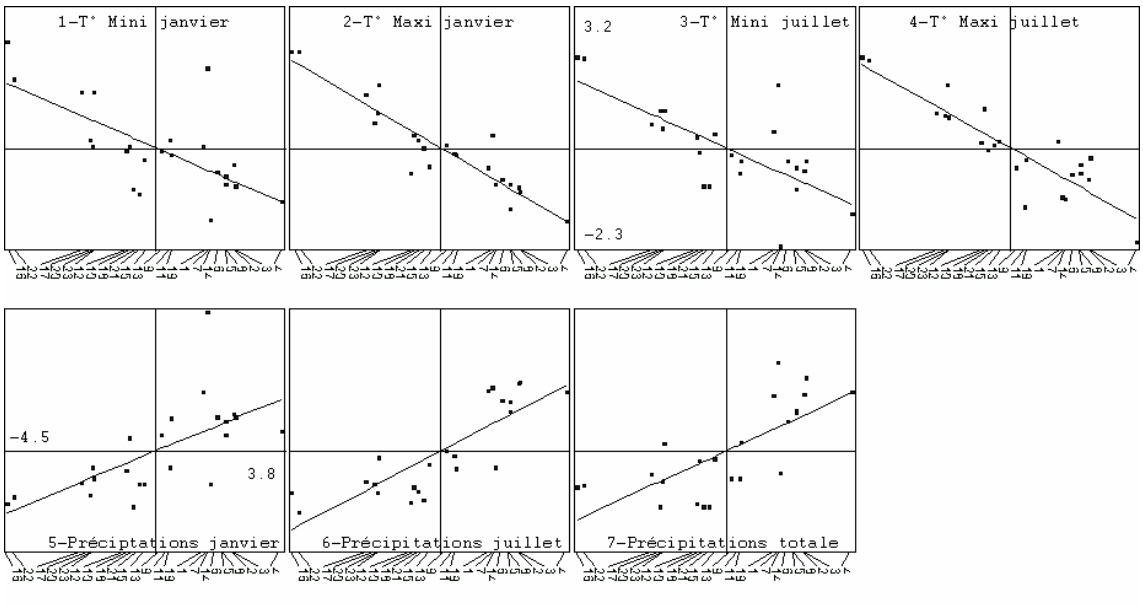
$$\mathbf{S}\mathbf{L} = \frac{1}{n}\mathbf{X}_*\mathbf{X}_*^t\mathbf{X}_*\mathbf{U} = \mathbf{X}_*\mathbf{R}\mathbf{U} = \mathbf{X}_*\mathbf{U}\Lambda = \mathbf{L}\Lambda$$

Les deux matrices **R** (*p*-*p*) et **S** (*n*-*n*) ont les mêmes valeurs propres non nulles (on démontre en outre que les autres valeurs propres de la plus grande sont toutes nulles). Les vecteurs que l'on cherche sont donc déjà connus. Ce sont, à une constante de normalisation près, les vecteurs dont les composantes sont les coordonnées des individus sur les axes principaux.

$L(\mathbf{y},\mathbf{X})$ ne peut pas dépasser λ_1 et l'atteint pour la variable normée $\frac{1}{\sqrt{\lambda_1}}\mathbf{l}_1$. On appelle ce vecteur la première composante principale (qui a donné son nom aux analyses). C'est la variable qui est la plus corrélée avec toutes les variables du tableau, d'où la figure dite graphe canonique.

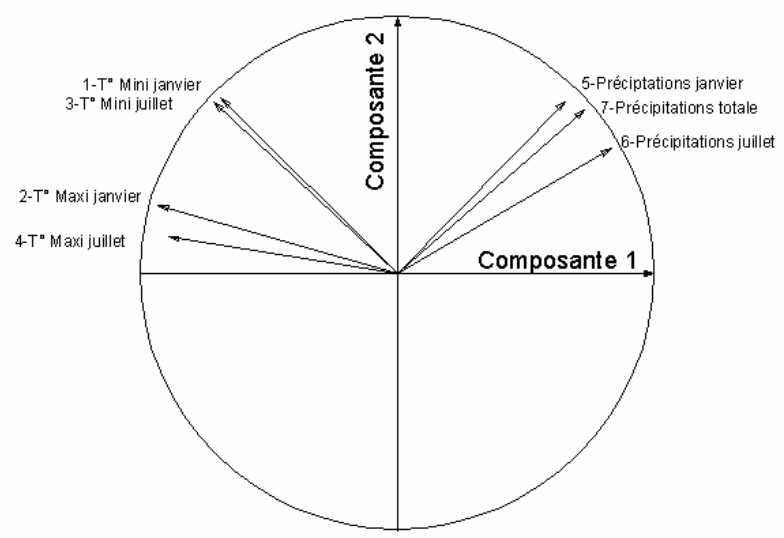
Quand on projette les variables sur les composantes principales on obtient les coordonnées des variables :

$$\mathbf{K} = \frac{1}{n}\mathbf{X}_*^t\mathbf{L}\Lambda^{-\frac{1}{2}} = \frac{1}{n}\mathbf{X}_*^t\mathbf{X}_*\mathbf{U}\Lambda^{-\frac{1}{2}} = \mathbf{R}\mathbf{U}\Lambda^{-\frac{1}{2}} = \mathbf{U}\Lambda^{\frac{1}{2}}$$



Graphe canonique en ACP normée

Comme les variables sont à une distance de 1 de l'origine, elles sont sur la sphère unitaire et se projette dans un cercle. Comme les coordonnées sont les corrélations entre variables et composantes, on appelle ces cartes des cercles de corrélation.



Cercle de corrélation en ACP normée

Ces éléments sont cohérents entre eux et l'interprétation des données utilisent les uns ou les autres suivant les cas concrets.