Statistique multivariée et Introduction à R

Serge Lhomme

Maître de conférences en Géographie
http://sergelhomme.fr/
serge.lhomme@u-pec.fr

1. Introduction
2. Corrélation et régression de variables quantitatives
3. Variables qualitatives
4. Méthodes exploratoires, synthétiques et classification
5. Interprétation des résultats obtenus sous R
6. Introduction à R

1 Introduction

2 Corrélation et régression de variables quantitatives

3 Variables qualitatives

4 Méthodes exploratoires, synthétiques et classification

5 Interprétation des résultats obtenus sous R

6 Introduction à R

Introduction

Les statistiques multivariées

Définition

En statistiques, les analyses multivariées ont pour caractéristique de s'intéresser à la distribution conjointe de plusieurs variables. Les analyses bivariées sont des cas particuliers à deux variables.

Les analyses multivariées sont très diverses selon l'objectif recherché ou la nature des variables. On peut identifier deux grandes familles :

- celle des méthodes descriptives visant à structurer et résumer l'information ;
- celle des méthodes explicatives visant à expliquer une ou des variables dites "dépendantes" (variables à expliquer) par un ensemble de variables dites "indépendantes" (variables explicatives).

1 Introduction

2 Corrélation et régression de variables quantitatives

3 Variables qualitatives

4 Méthodes exploratoires, synthétiques et classification

5 Interprétation des résultats obtenus sous R

6 Introduction à R

Corrélation et régression

Définitions

Corrélation

Etudier la corrélation entre deux ou plusieurs variables, c'est mesurer l'intensité de la liaison qui peut exister entre ces variables.

Régression

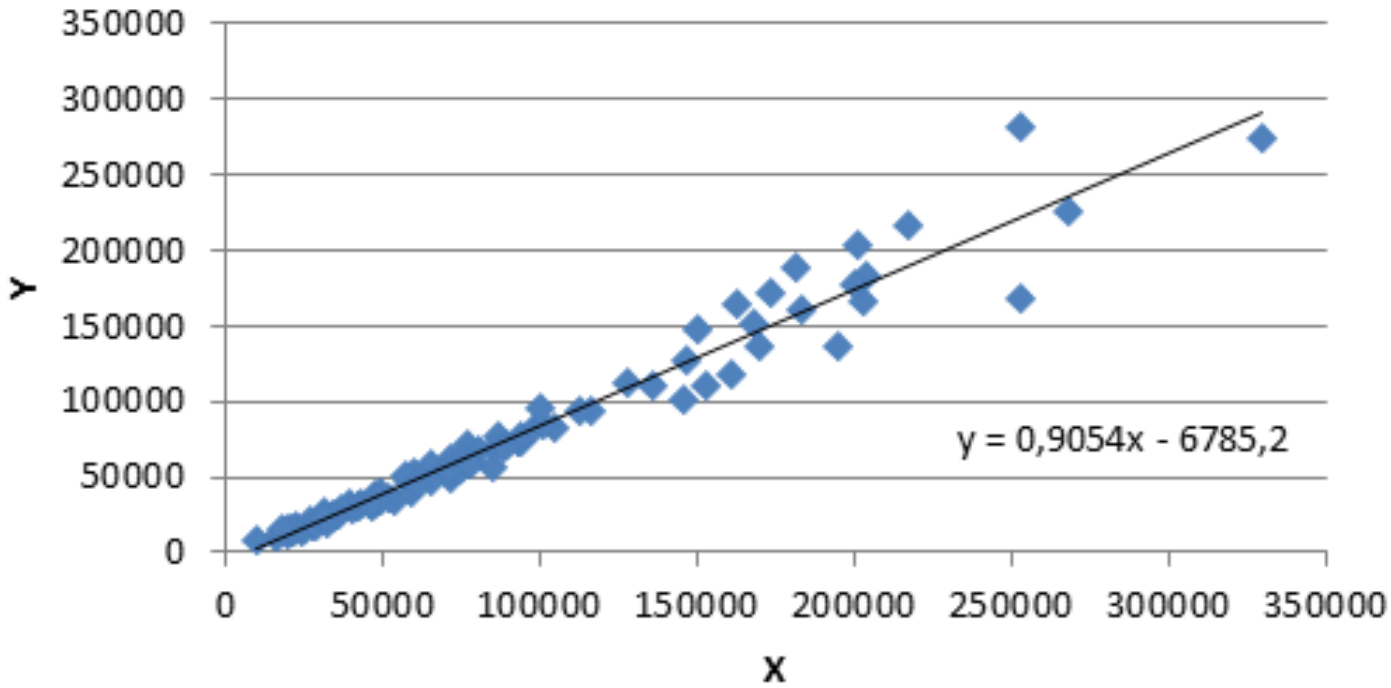
La régression est un ensemble de méthodes statistiques très utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres. On cherche alors à retrouver (prédire) la variable à expliquer à l'aide des autres variables.

Dans le cadre de corrélations et de régressions linéaires, on s'intéresse plus particulièrement à des relations de type linéaire.

Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

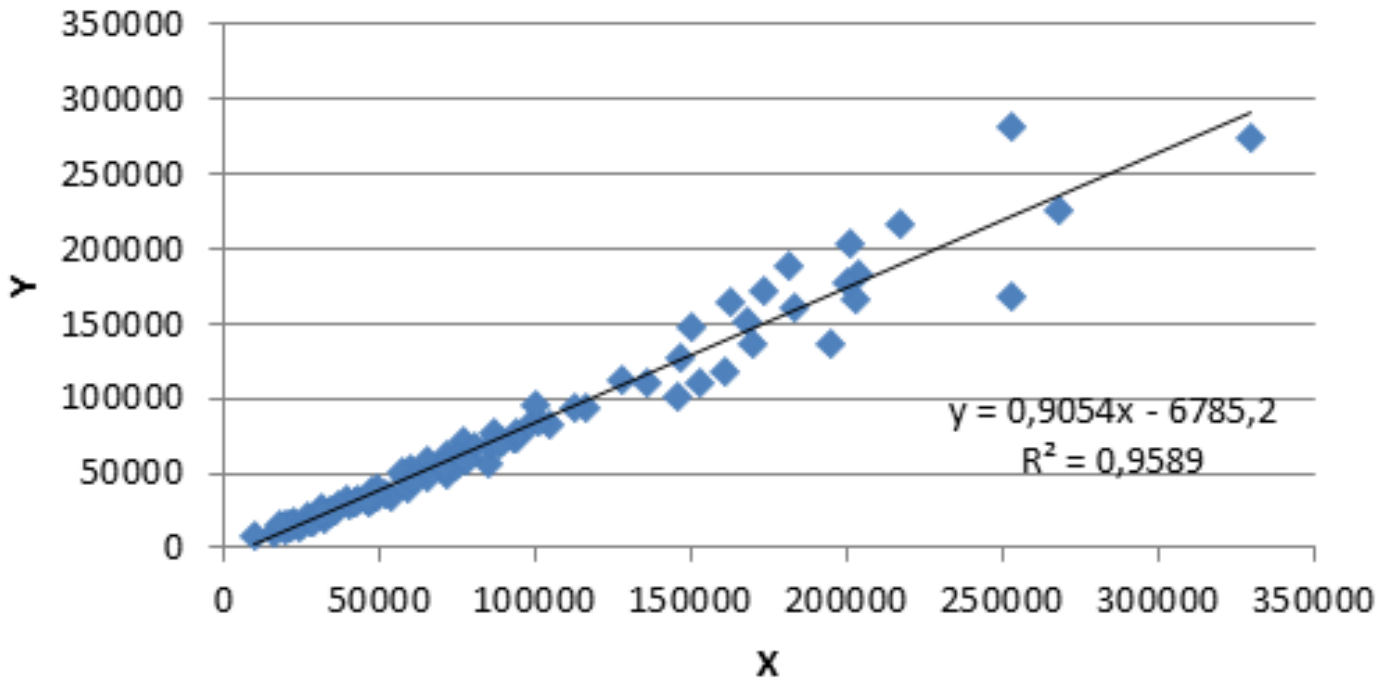
Pour faire simple, lorsque l'on étudie deux variables quantitatives, on peut produire un « nuage de points », une régression linéaire vise alors à résumer ce nuage de points par une forme plus simple à interpréter indiquant la « tendance générale » : une droite.



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

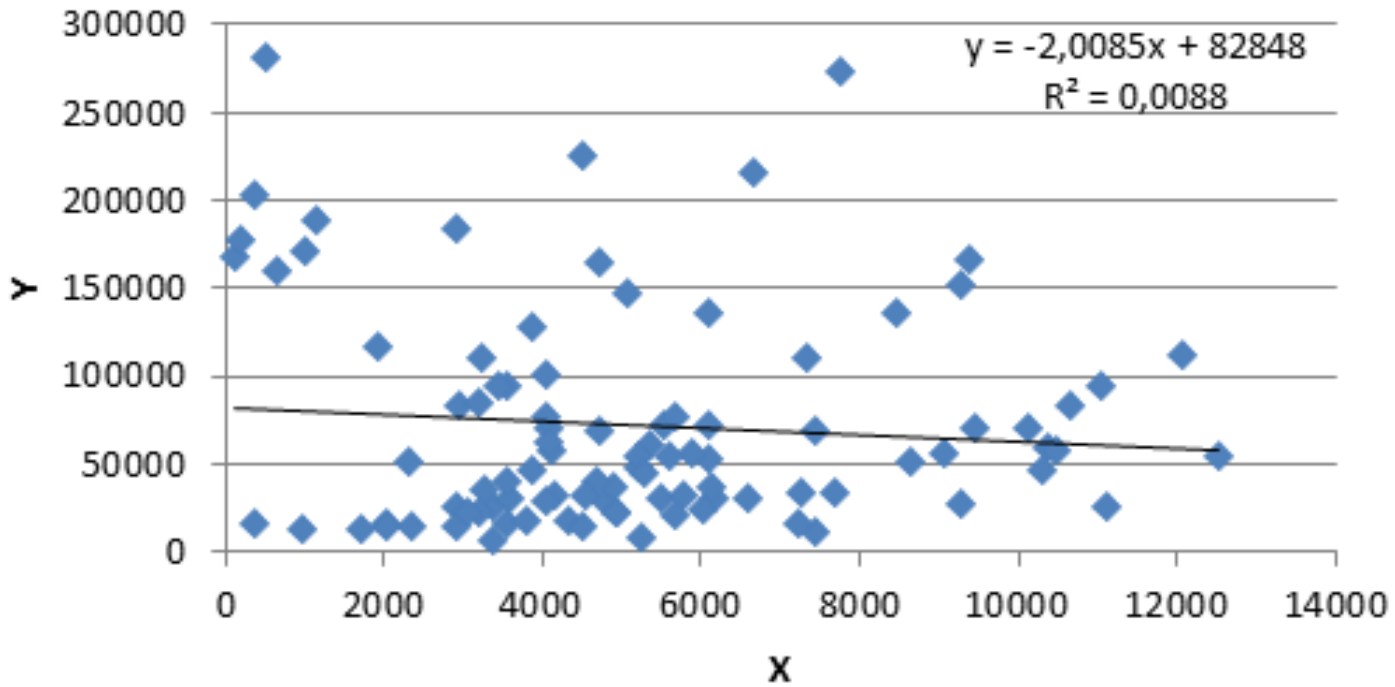
C'est le coefficient de corrélation (R) ou le coefficient de détermination (R^2) qui nous permet de dire si cette régression est « juste », à quel point la droite résume bien les variations du nuage de points :



Corrélation et régression linéaire

Des graphiques plutôt que des définitions ou des calculs pour comprendre

C'est le coefficient de corrélation ou le coefficient de détermination qui nous permet de dire si cette régression est « juste », ou « pas du tout » :



Corrélation et régression linéaire

Equation

L'équation de la droite recherchée est de type linéaire :

$$Y = aX + b$$

Les coefficients de la droite de régression s'obtiennent alors avec les équations suivantes :

$$a = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \text{ et } b = \bar{y} - a\bar{x}$$

Le coefficient de corrélation R (de Bravais-Pearson) :

$$R = Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y} \text{ (-1} \leq \text{R} \leq \text{1)}$$

Corrélation et régression linéaire

Les résidus

Définition

Un résidu est dans une régression le terme qui n'est pas expliqué par la ou les variables explicatives.

Il se calcule simplement en calculant l'écart entre la valeur réelle de y et la valeur théorique (estimée, prédite) de y (obtenue à partir de l'équation déterminée par la régression linéaire) :

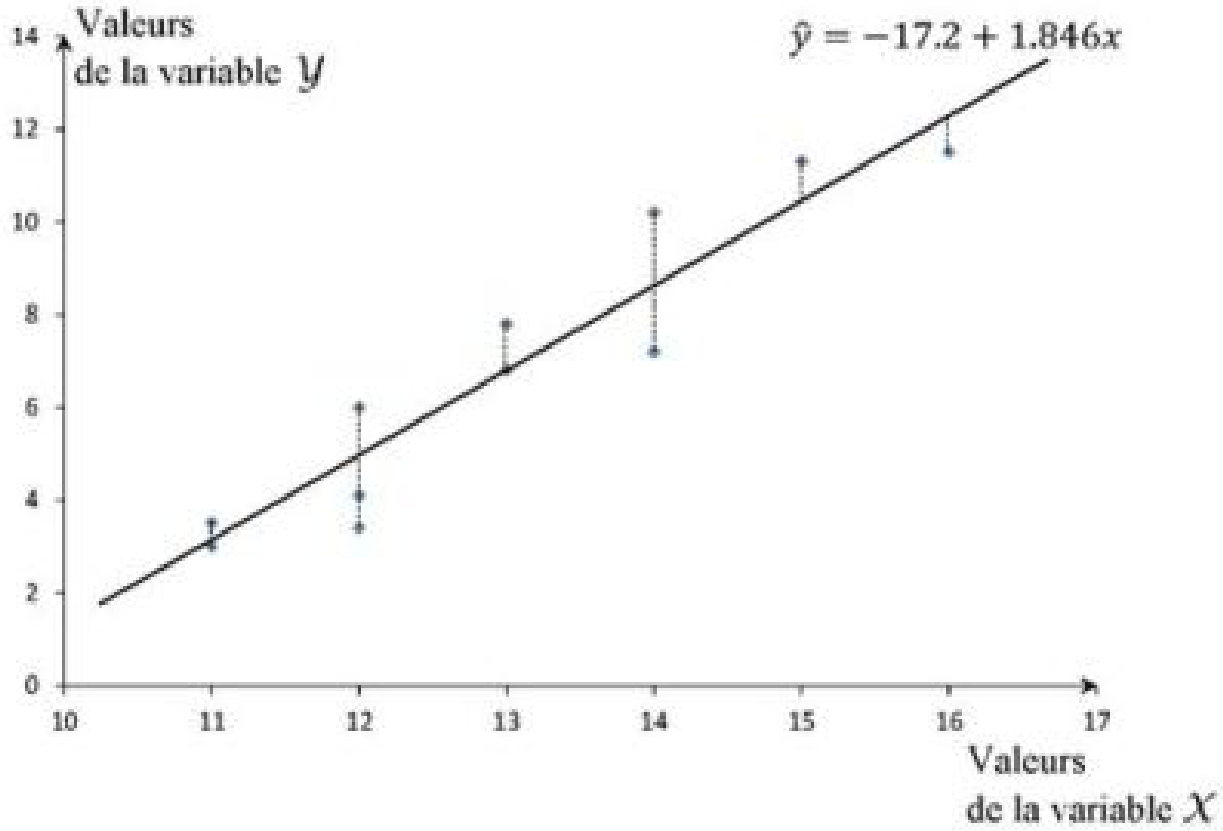
$$\hat{Y}i = aXi + b$$

$$e_i = Yi - \hat{Y}i$$

$$Yi = \hat{Y}i + e_i = aXi + b + e_i$$

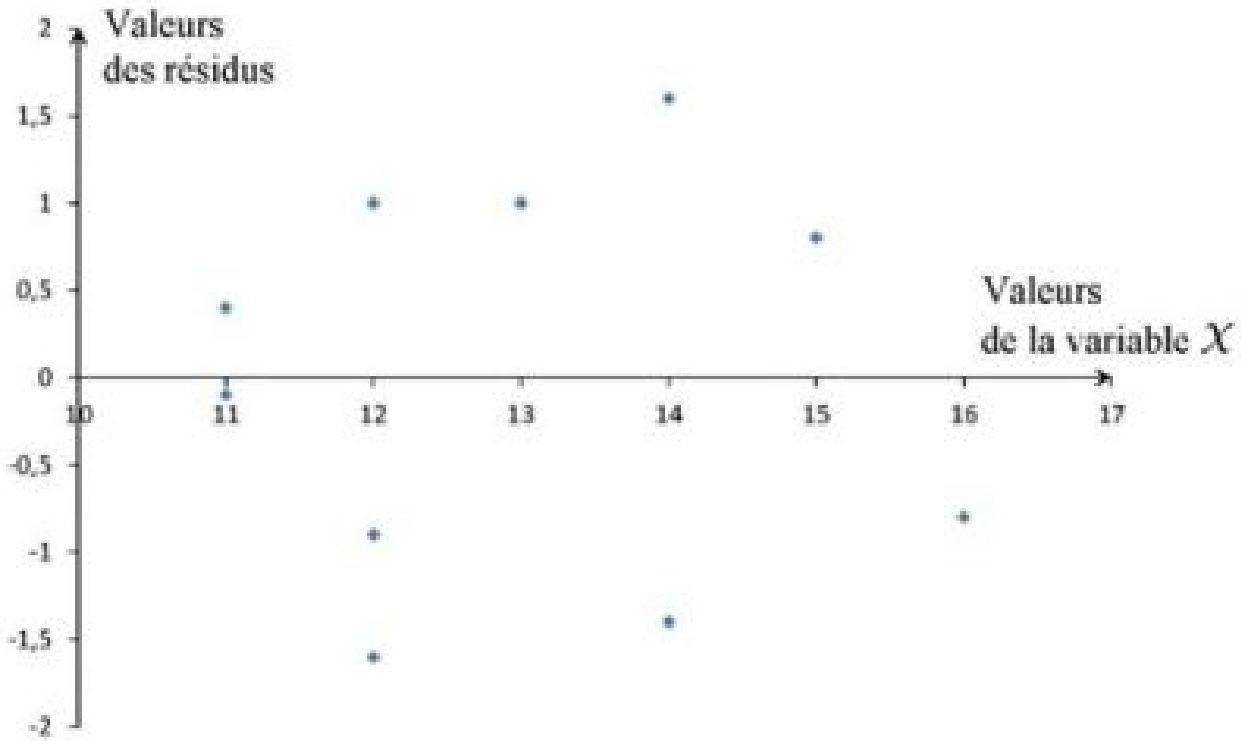
Corrélation et régression linéaire

Les résidus



Corrélation et régression linéaire

Les résidus



Corrélation et régression linéaire

MCO : Méthode des moindres carrées ordinaires

Selon la méthode des moindres carrées (de Legendre et Gauss), la droite qui décrit « le mieux » les données est celle qui minimise la somme quadratique des résidus (des déviations des mesures aux prédictions).

$$y_i - \bar{y} = \widehat{y}_i - \bar{y} + e_i$$

$$\sum_i^n (y_i - \bar{y})^2 = \sum_i^n (\widehat{y}_i - \bar{y})^2 + \sum_i^n {e_i}^2$$

$$SCT = SCE + SCR$$

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Ainsi, R^2 peut être interprété comme la part (le pourcentage) de variance expliquée par les valeurs obtenues par la droite de régression ($0 \leq R^2 \leq 1$).

La significativité

Test T de Student

Dans les faits, il est important de savoir si les coefficients calculés sont significativement différents de ce que l'on pourrait obtenir par hasard entre deux variables aléatoires de même taille.

Pour le coefficient de corrélation, il faut comparer la valeur t obtenue avec celle du tableau de Student pour n - 2 degrés de liberté :

$$t = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.9975	0.9995
1	0.1584	0.3249	0.5095	0.7265	1	1.3764	1.9626	3.0777	6.3137	12.706	31.821	63.656	127.32	636.58
2	0.1421	0.2887	0.4447	0.6172	0.8165	1.0607	1.3862	1.8856	2.92	4.3027	6.9645	9.925	14.089	31.6
3	0.1366	0.2767	0.4242	0.5844	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8408	7.4532	12.924
4	0.1338	0.2707	0.4142	0.5686	0.7407	0.941	1.1896	1.5332	2.1318	2.7765	3.7469	4.6041	5.5975	8.6101
5	0.1322	0.2672	0.4082	0.5594	0.7267	0.9195	1.1558	1.4759	2.015	2.5706	3.3649	4.0321	4.7733	6.8685
6	0.1311	0.2648	0.4043	0.5534	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168	5.9587
7	0.1303	0.2632	0.4015	0.5491	0.7111	0.896	1.1192	1.4149	1.8946	2.3646	2.9979	3.4995	4.0294	5.4081
8	0.1297	0.2619	0.3995	0.5459	0.7064	0.8889	1.1081	1.3968	1.8595	2.306	2.8965	3.3554	3.8325	5.0414
9	0.1293	0.261	0.3979	0.5435	0.7027	0.8834	1.0997	1.383	1.8331	2.2622	2.8214	3.2498	3.6896	4.7809
10	0.1289	0.2602	0.3966	0.5415	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.5868
11	0.1286	0.2596	0.3956	0.5399	0.6974	0.8755	1.0877	1.3634	1.7959	2.201	2.7181	3.1058	3.4966	4.4369
12	0.1283	0.259	0.3947	0.5386	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.681	3.0545	3.4284	4.3178

La significativité

La p-value

	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95	0.975	0.99	0.995	0.9975	0.9995
1	0.1584	0.3249	0.5095	0.7265	1	1.3764	1.9626	3.0777	6.3137	12.706	31.821	63.656	127.32	636.58
2	0.1421	0.2887	0.4447	0.6172	0.8165	1.0607	1.3862	1.8856	2.92	4.3027	6.9645	9.925	14.089	31.6
3	0.1366	0.2767	0.4242	0.5844	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8408	7.4532	12.924
4	0.1338	0.2707	0.4142	0.5686	0.7407	0.941	1.1896	1.5332	2.1318	2.7765	3.7469	4.6041	5.5975	8.6101
5	0.1322	0.2672	0.4082	0.5594	0.7267	0.9195	1.1558	1.4759	2.015	2.5706	3.3649	4.0321	4.7733	6.8685
6	0.1311	0.2648	0.4043	0.5534	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074	4.3168	5.9587
7	0.1303	0.2632	0.4015	0.5491	0.7111	0.896	1.1192	1.4149	1.8946	2.3646	2.9979	3.4995	4.0294	5.4081
8	0.1297	0.2619	0.3995	0.5459	0.7064	0.8889	1.1081	1.3968	1.8595	2.306	2.8965	3.3554	3.8325	5.0414
9	0.1293	0.261	0.3979	0.5435	0.7027	0.8834	1.0997	1.383	1.8331	2.2622	2.8214	3.2498	3.6896	4.7809
10	0.1289	0.2602	0.3966	0.5415	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693	3.5814	4.5868
11	0.1286	0.2596	0.3956	0.5399	0.6974	0.8755	1.0877	1.3634	1.7959	2.201	2.7181	3.1058	3.4966	4.4369
12	0.1283	0.259	0.3947	0.5386	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.681	3.0545	3.4284	4.3178

40	0.1265	0.255	0.3881	0.5286	0.6807	0.8507	1.05	1.3031	1.6839	2.0211	2.4233	2.7045	2.9712	3.551
41	0.1264	0.255	0.388	0.5285	0.6805	0.8505	1.0497	1.3025	1.6829	2.0195	2.4208	2.7012	2.967	3.5443
42	0.1264	0.255	0.388	0.5284	0.6804	0.8503	1.0494	1.302	1.682	2.0181	2.4185	2.6981	2.963	3.5377
43	0.1264	0.2549	0.3879	0.5283	0.6802	0.8501	1.0491	1.3016	1.6811	2.0167	2.4163	2.6951	2.9592	3.5316
44	0.1264	0.2549	0.3878	0.5282	0.6801	0.8499	1.0488	1.3011	1.6802	2.0154	2.4141	2.6923	2.9555	3.5258
45	0.1264	0.2549	0.3878	0.5281	0.68	0.8497	1.0485	1.3007	1.6794	2.0141	2.4121	2.6896	2.9521	3.5203
46	0.1264	0.2548	0.3877	0.5281	0.6799	0.8495	1.0482	1.3002	1.6787	2.0129	2.4102	2.687	2.9488	3.5149
47	0.1263	0.2548	0.3877	0.528	0.6797	0.8493	1.048	1.2998	1.6779	2.0117	2.4083	2.6846	2.9456	3.5099
48	0.1263	0.2548	0.3876	0.5279	0.6796	0.8492	1.0478	1.2994	1.6772	2.0106	2.4066	2.6822	2.9426	3.505
49	0.1263	0.2547	0.3876	0.5278	0.6795	0.849	1.0475	1.2991	1.6766	2.0096	2.4049	2.68	2.9397	3.5005
50	0.1263	0.2547	0.3875	0.5278	0.6794	0.8489	1.0473	1.2987	1.6759	2.0086	2.4033	2.6778	2.937	3.496
60	0.1262	0.2545	0.3872	0.5272	0.6786	0.8477	1.0455	1.2958	1.6706	2.0003	2.3901	2.6603	2.9146	3.4602
70	0.1261	0.2543	0.3869	0.5268	0.678	0.8468	1.0442	1.2938	1.6669	1.9944	2.3808	2.6479	2.8987	3.435
80	0.1261	0.2542	0.3867	0.5265	0.6776	0.8461	1.0432	1.2922	1.6641	1.9901	2.3739	2.6387	2.887	3.4164
90	0.126	0.2541	0.3866	0.5263	0.6772	0.8456	1.0424	1.291	1.662	1.9867	2.3685	2.6316	2.8779	3.4019
100	0.126	0.254	0.3864	0.5261	0.677	0.8452	1.0418	1.2901	1.6602	1.984	2.3642	2.6259	2.8707	3.3905
110	0.126	0.254	0.3863	0.5259	0.6767	0.8449	1.0413	1.2893	1.6588	1.9818	2.3607	2.6213	2.8648	3.3811
120	0.1259	0.2539	0.3862	0.5258	0.6765	0.8446	1.0409	1.2886	1.6576	1.9799	2.3578	2.6174	2.8599	3.3734
130	0.1259	0.2539	0.3862	0.5257	0.6764	0.8444	1.0406	1.2881	1.6567	1.9784	2.3554	2.6142	2.8557	3.367
140	0.1259	0.2538	0.3861	0.5256	0.6762	0.8442	1.0403	1.2876	1.6558	1.9771	2.3533	2.6114	2.8522	3.3613
infini (loi normale)	0.1257	0.2533	0.3853	0.5244	0.6744	0.8416	1.0364	1.2816	1.6449	1.96	2.3264	2.5759	2.8072	3.2908

Si $| t | > t_{seuil}$, on rejette l'hypothèse nulle pour le risque choisi (classiquement 5%).

La significativité

La p-value

Plutôt que de comparer la statistique calculée avec le seuil théorique fourni 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.

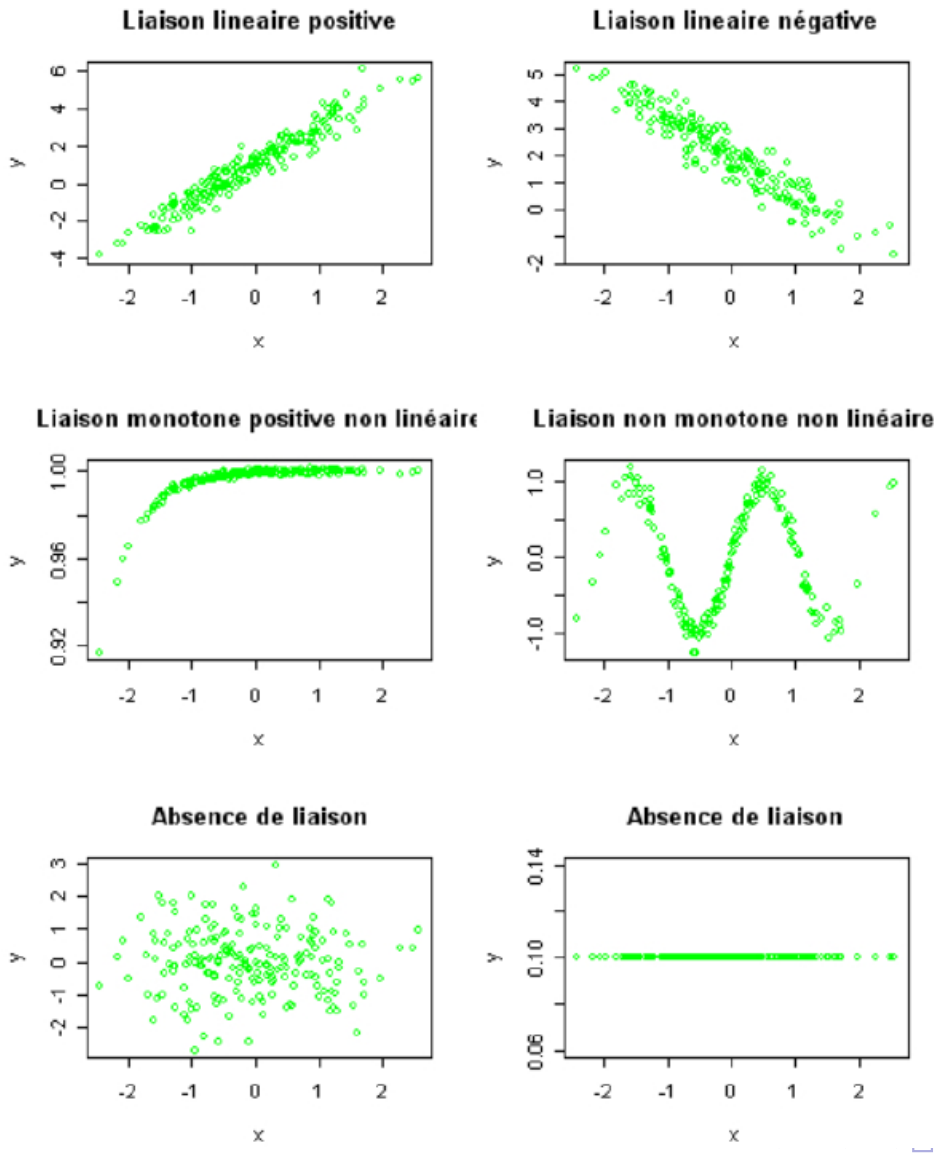
La probabilité critique correspond au niveau de risque à partir duquel on ne peut plus rejeter l'hypothèse nulle.

Très souvent la p-value est simplement interprétée comme la probabilité de se tromper en rejetant l'hypothèse nulle.

De même des tests de significativité (Student) peuvent être effectués sur les coefficients de la régression (les estimateurs) dont la robustesse (la qualité) dépend notamment de la variation des variables explicatives. Les erreurs types correspondantes permettent de déterminer des intervalles de confiance.

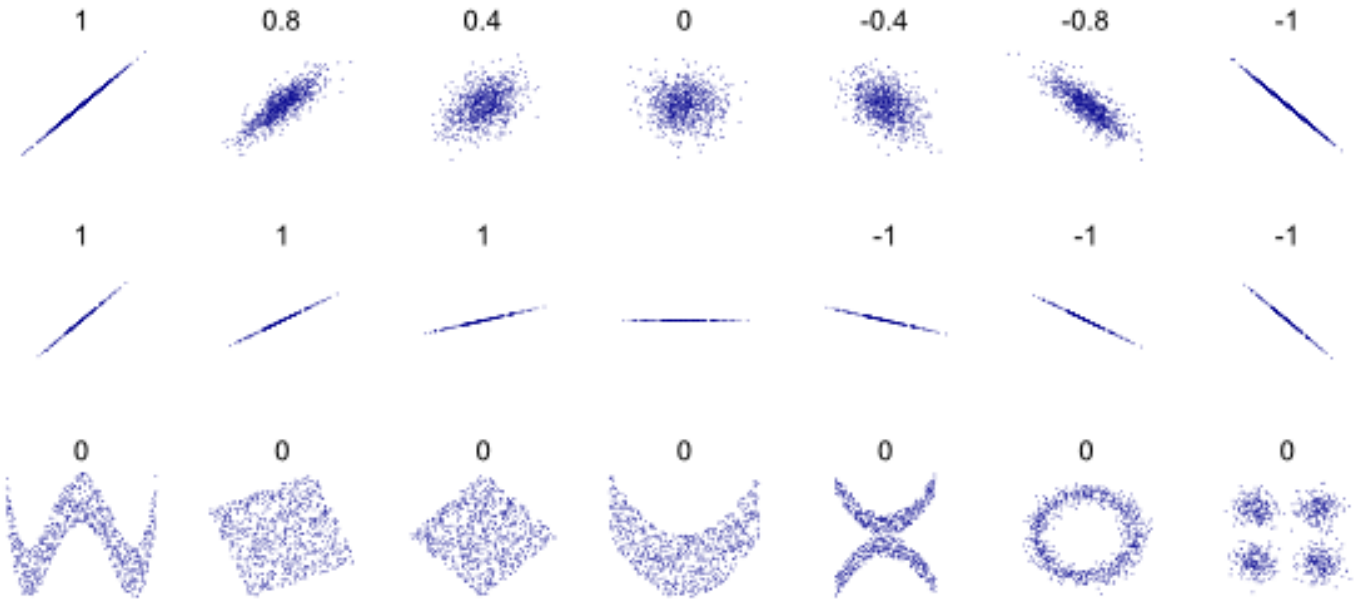
Régression non linéaire

Différentes formes de nuages de points



Régression non linéaire

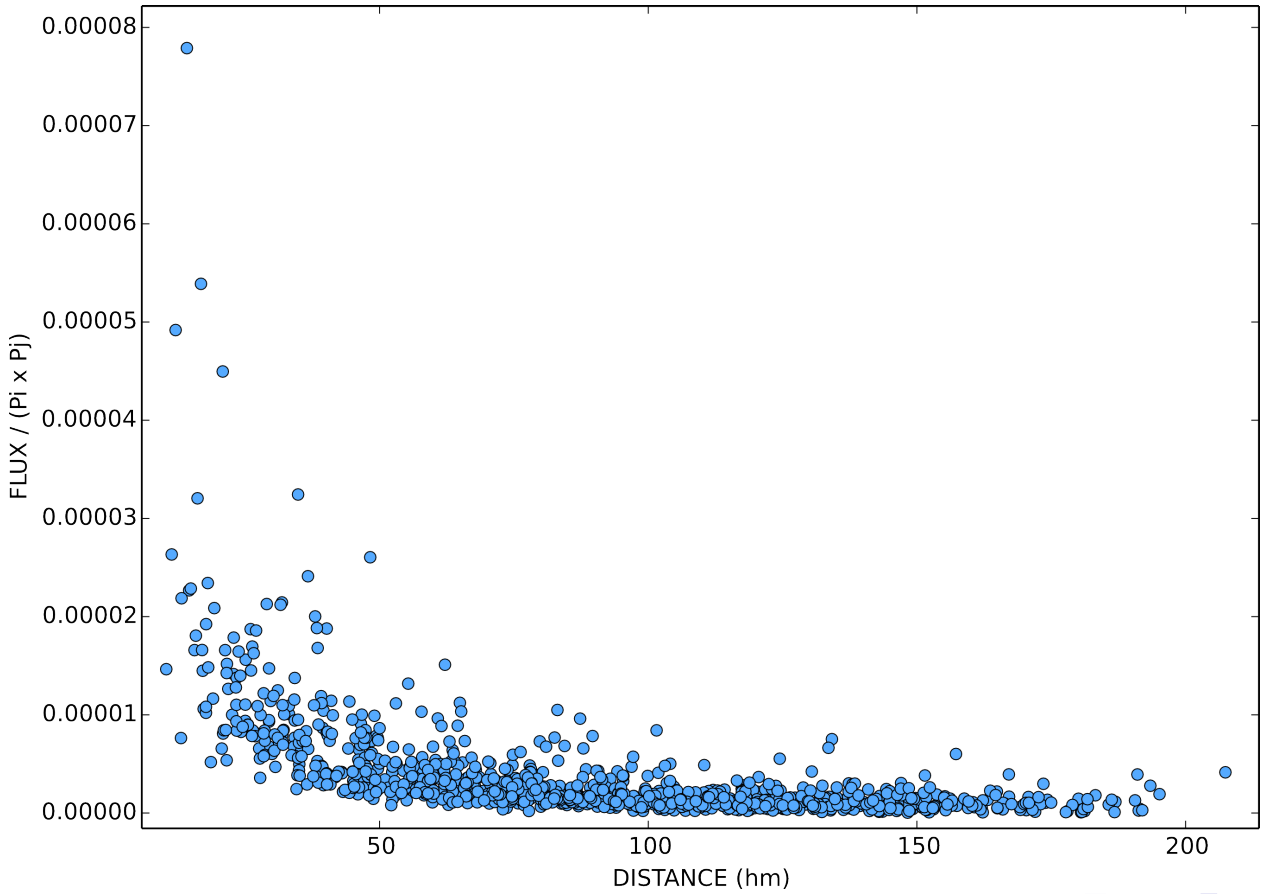
Différentes formes de nuages de points



Régression non linéaire

Kernel trick : Astuce du noyau

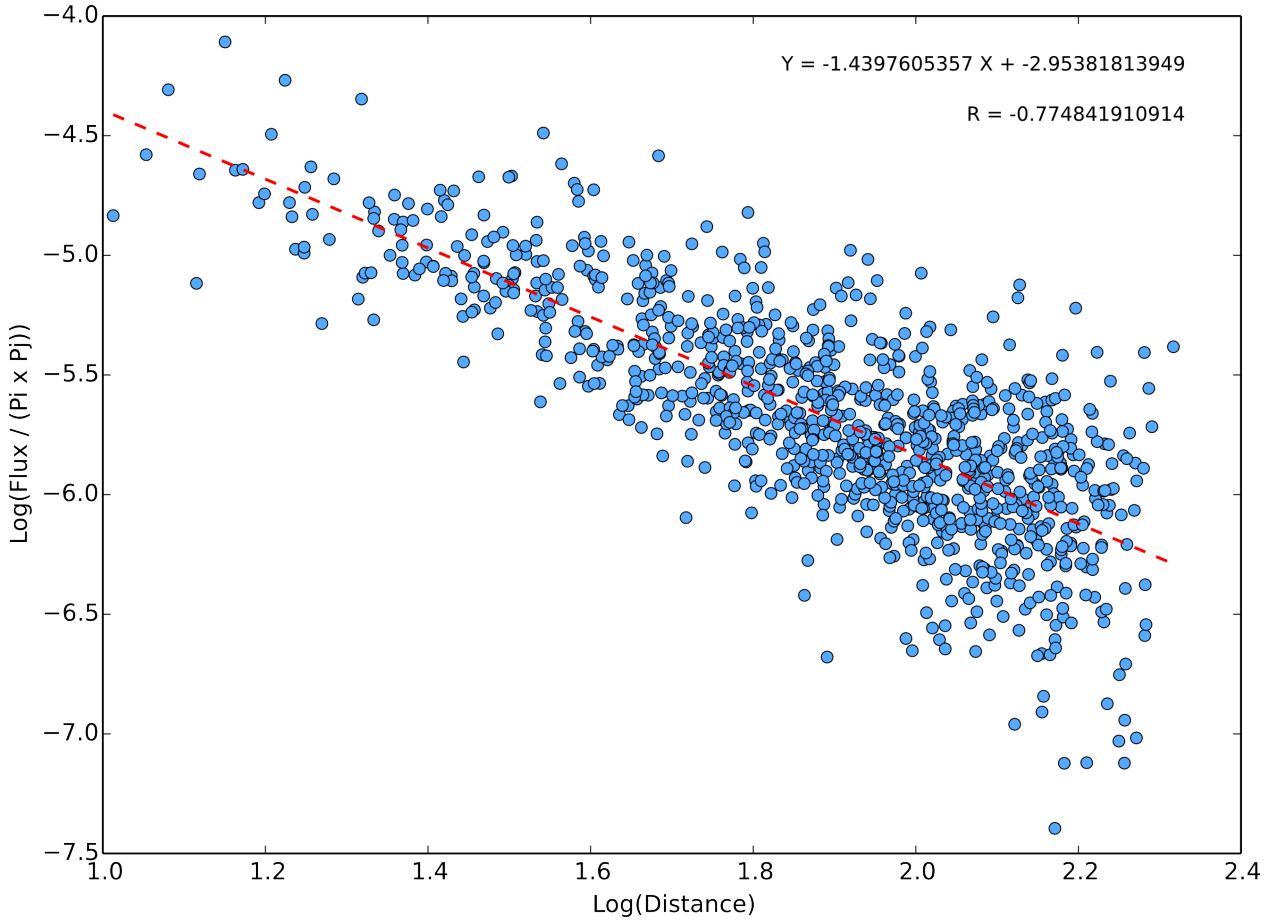
En apprentissage automatique, l'astuce du noyau (kernel tricks), est une méthode qui permet d'utiliser un classifieur linéaire pour résoudre un problème non linéaire.



Régression non linéaire

Kernel trick : Astuce du noyau

Pour des régressions, cela revient à utiliser des fonctions mathématiques non-linéaires pour revenir à une situation linéaire.



Régression non linéaire

Détails de la transformation bi-logarithmique

$$P = K \times n^{-\alpha}$$

$$\Rightarrow \log(P) = \log(K \times n^{-\alpha})$$

$$\Rightarrow \log(P) = \log(K) + \log(n^{-\alpha})$$

$$\Rightarrow \log(P) = \log(K) - \alpha \log(n)$$

$$\Rightarrow \log(P) = -\alpha \log(n) + \log(K)$$

$$\Rightarrow Y = -aX + b$$

Conclusion de la transformation bi-logarithmique

$$b = \log(K) \Rightarrow K = 10^b$$

$$a = \alpha$$

Régression non linéaire

Une astuce pas si simple à appliquer

Attention bidouille

Attention, le kernel trick est une "bidouille" et n'est pas totalement sans impact sur les calculs.

Cette bidouille est très utilisée et il ne faut pas la dramatiser. Certes les estimateurs pourront alors être biaisés, mais cette méthode reste très efficace.

Difficile

Il est souvent très difficile de trouver les transformations mathématiques à appliquer.

Régression linéaire multiple

Une simple généralisation d'une régression linéaire bivariée

Lorsque l'on cherche à expliquer un phénomène d'un point de vue statistique, il est rare qu'une seule variable explicative soit suffisante. Un « modèle »nécessite ainsi souvent l'intégration de plusieurs variables.

Heureusement les méthodes de régression linéaire se généralisent très bien à plusieurs variables. On parle de régression linéaire multiple. On tente alors d'expliquer une variable Y par plusieurs variables explicatives Xi.

Si vous utilisez deux variables explicatives, votre nuage de points pourra être représenté en trois dimensions et les méthodes de régression vous proposeront alors l'équation d'un plan pour représenter ce nuage...

Le calcul du R^2 est strictement identique et donne en quelque sorte la précision du modèle, la part de variance expliquée de la variable Y par la combinaison linéaire des variables explicatives Xi.

$$Y = a1X1 + a2X2 + a3X3 + ... + b$$

Régression linéaire multiple

Une simple généralisation, mais quelques difficultés

La notion de corrélation reste « bivariée » et ne se généralise pas, mais les coefficients des variables explicatives peuvent faire l'objet de tests de significativité.

Il faut alors faire attention à la colinéarité entre les variables, c'est à dire aux éventuelles corrélations entre les variables explicatives. Ainsi, en théorie, une régression linéaire multiple doit s'accompagner d'études de corrélation par paire (matrice des corrélations).

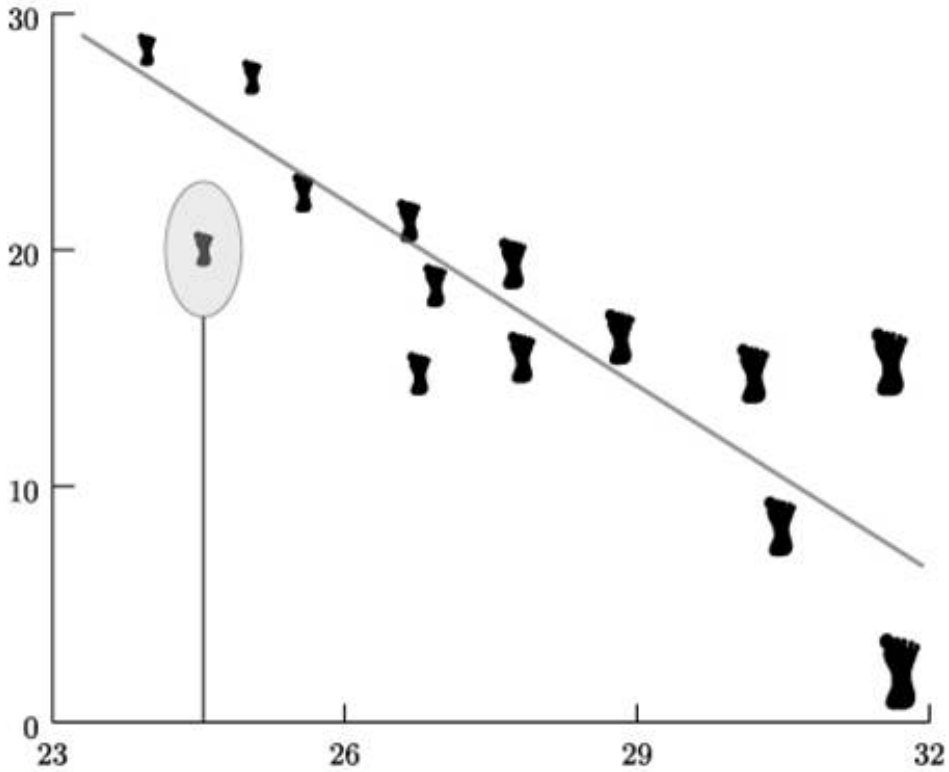
Ajouter des variables, même non significatives, contribue à améliorer le modèle explicatif en termes de R^2. C'est pourquoi un modèle doit viser à rester parcimonieux, en gardant uniquement des variables significatives.

Il existe aussi des alternatives au calcul du R^2, comme le critère d'information d'Akaike ou le critère d'information bayésien.

Les pièges à éviter

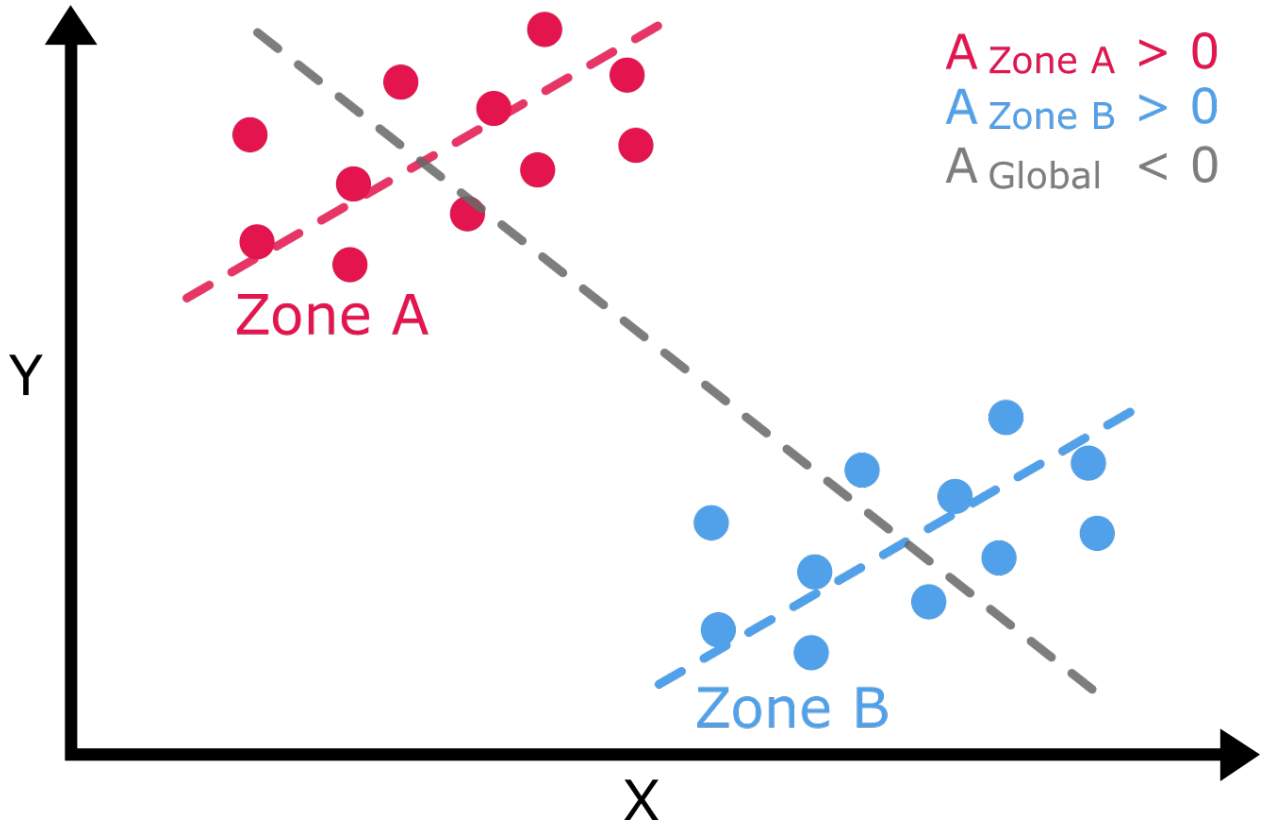
Corrélation n'est pas causalité

Nombre de fautes d'orthographe en fonction de la pointure. Les élèves ayant les plus grands pieds font moins de fautes...



Les pièges à éviter

Paradoxe de Simpson et facteur confondant



Les pièges à éviter

L'erreur écologique

En géographie, l'étude des corrélations se fait à travers l'analyse d'un ensemble de lieux, de territoires au sein desquels on a des agrégats.

Ainsi lorsque les variables décrivant ces lieux sont des attributs sociaux décrivant des habitants, il faut toujours faire attention au fait qu'une corrélation établie au niveau des lieux n'implique pas forcément une corrélation au niveau des individus.

Une étude menée au niveau des individus (sociologique) peut montrer que le taux de criminalité est plus élevé chez les autochtones que chez les étrangers. Pourtant, cette étude au niveau des quartiers (géographique) peut très bien montrer une corrélation entre la proportion d'étrangers des quartiers et leur taux de criminalité.

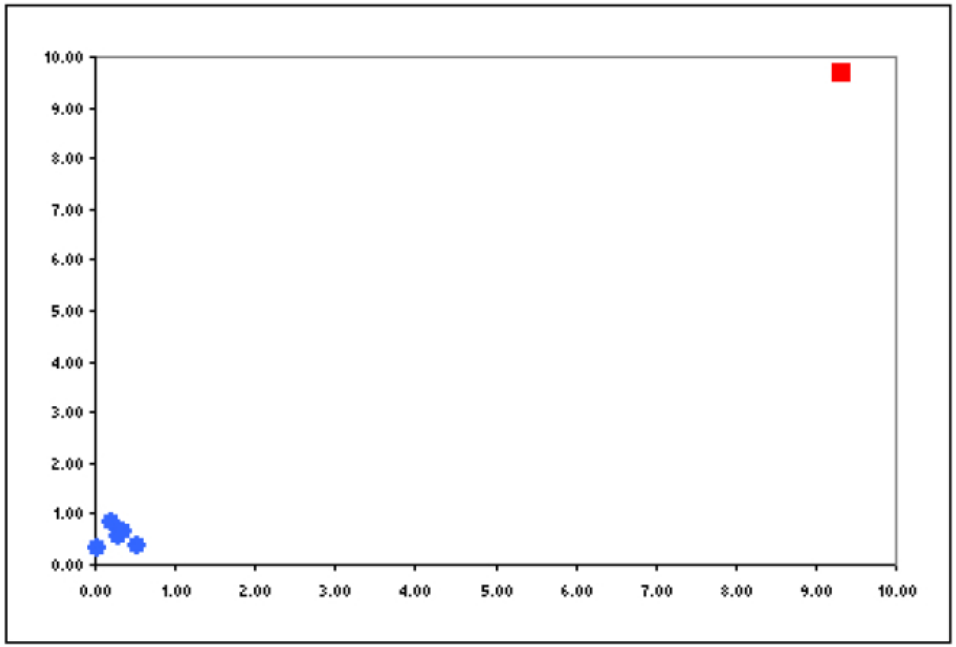
Il faut faire attention à ne pas « individualiser » une corrélation issue d'un « agrégat ».

Les pièges à éviter

Les points atypiques aberrants

	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**



Les pièges à éviter

Les règles statistiques à respecter

Régression et corrélation s'appliquent en théorie uniquement à des variables quantitatives continues.

Le test de Student classiquement associé au calcul de régression dans les logiciels statistiques s'applique à des distributions normales.

Le test de Shapiro-Wilk permet de tester la normalité d'une distribution.

Les tests de corrélation de Kendall et de Spearman sont recommandés lorsque les variables ne suivent pas une loi normale.

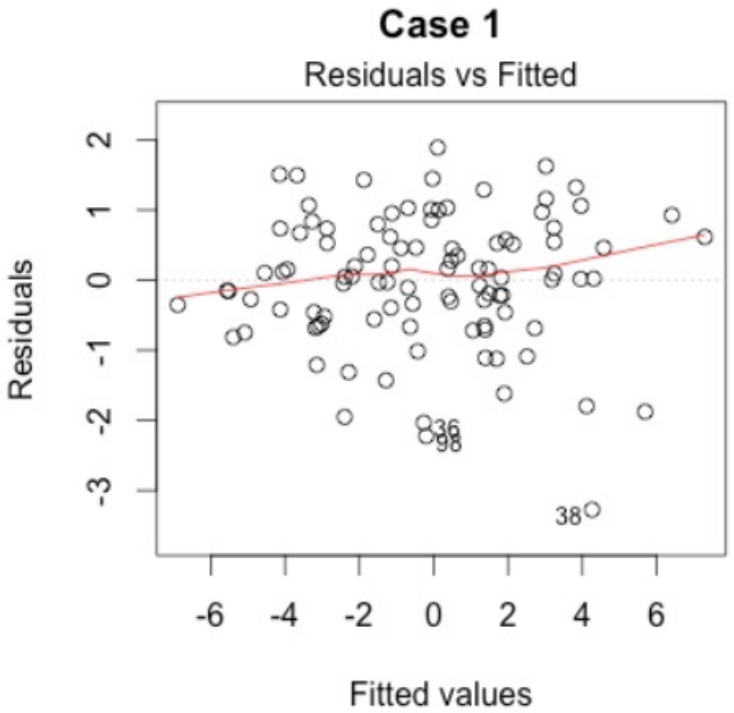
Les statistiques τ (tau) de kendall et ρ (rho) de Spearman sont respectivement utilisées pour estimer des coefficients de corrélation fondés sur les rangs.

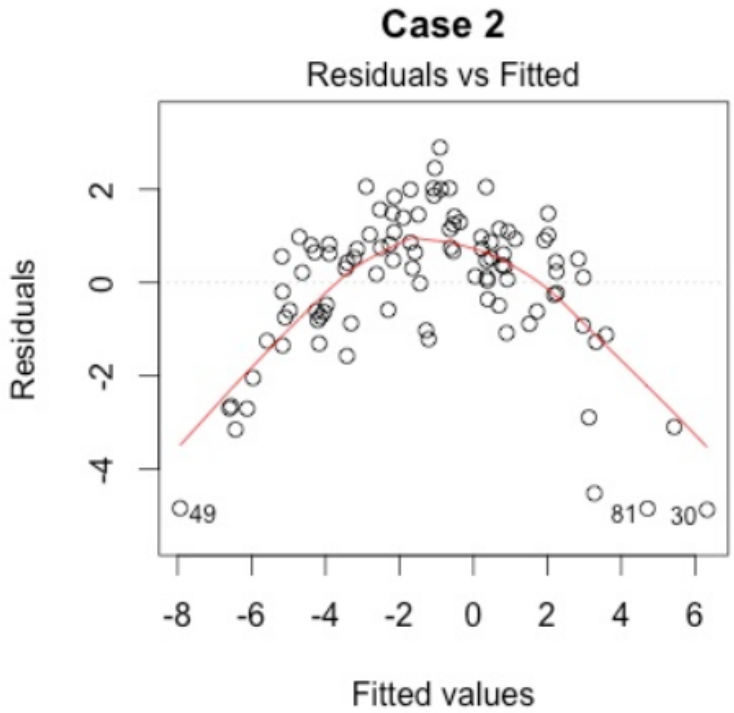
Ce sont des tests statistiques dits robustes, car ils ne dépendent pas de la distribution des données. Ils sont non paramétriques.

Les pièges à éviter

Etudier différents graphiques des résidus pour éviter des erreurs

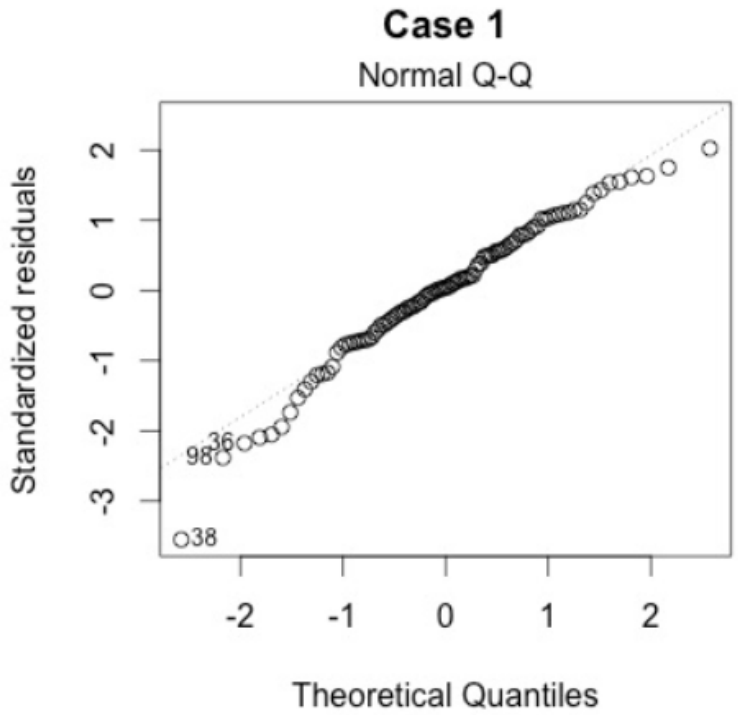
Hypothèses nombreuses à vérifier sur les erreurs : Distribution normale des erreurs ; Indépendance des erreurs (attention à l'autocorrélation temporelle ou spatiale) ; Exogénéité (variables explicatives non corrélées au terme d'erreur) ; Homoscédasticité (les termes d'erreurs sont supposés de variance constante).

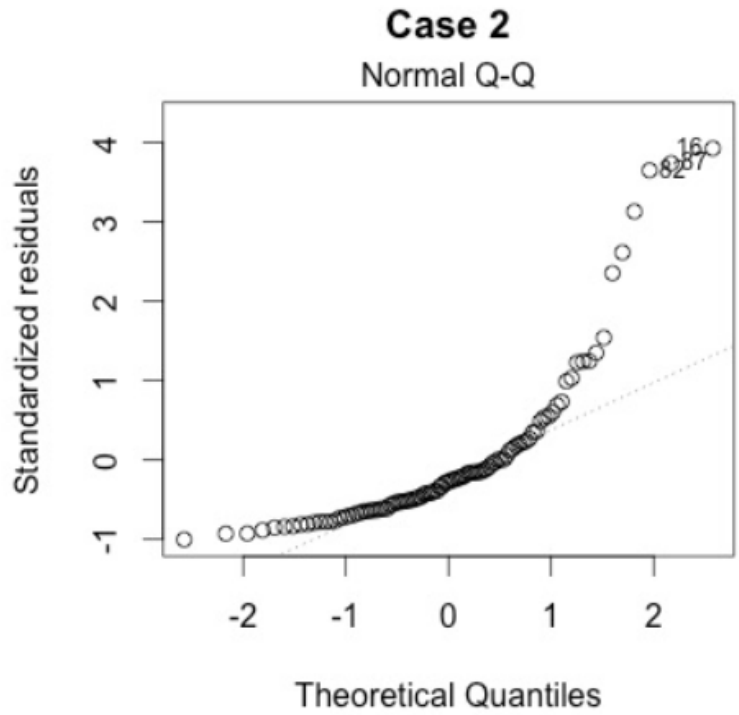




Les pièges à éviter

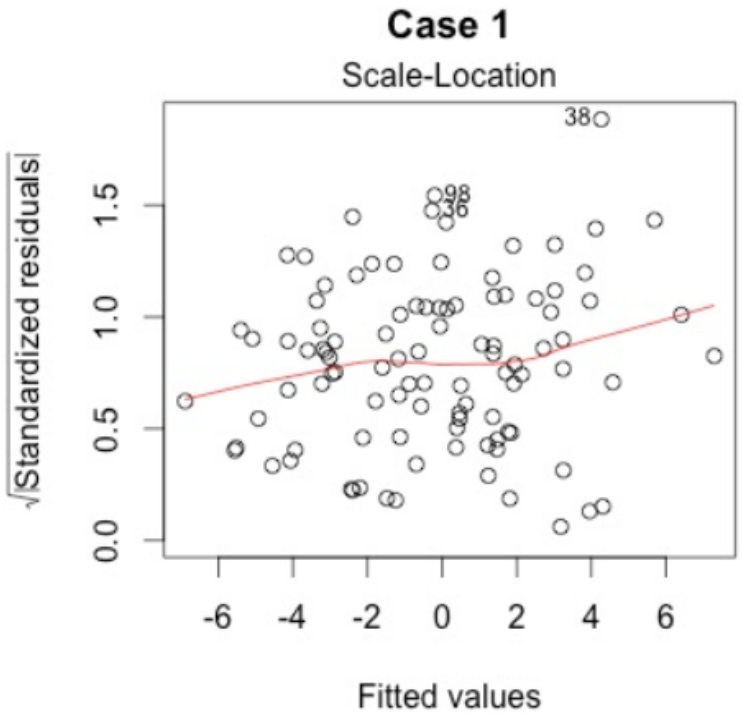
Etudier différents graphiques des résidus pour éviter des erreurs

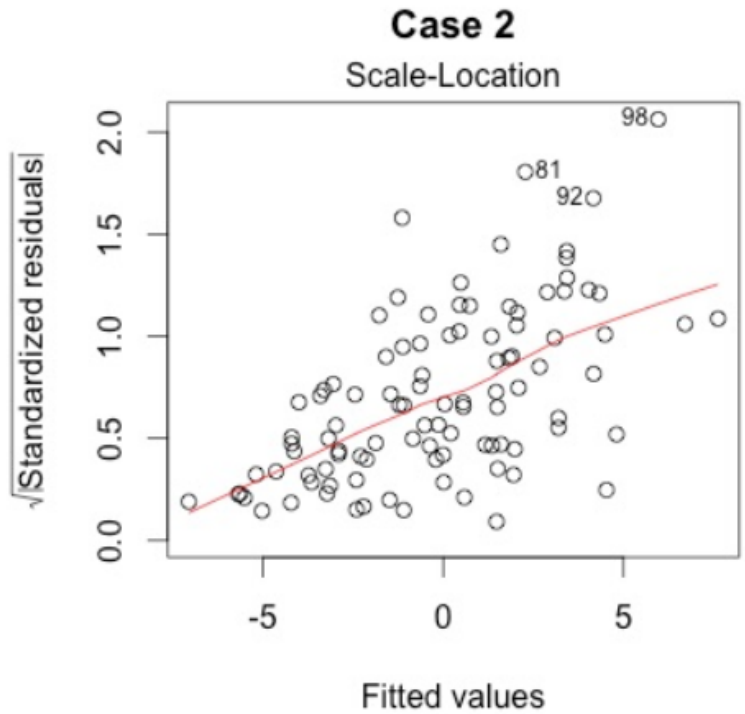




Les pièges à éviter

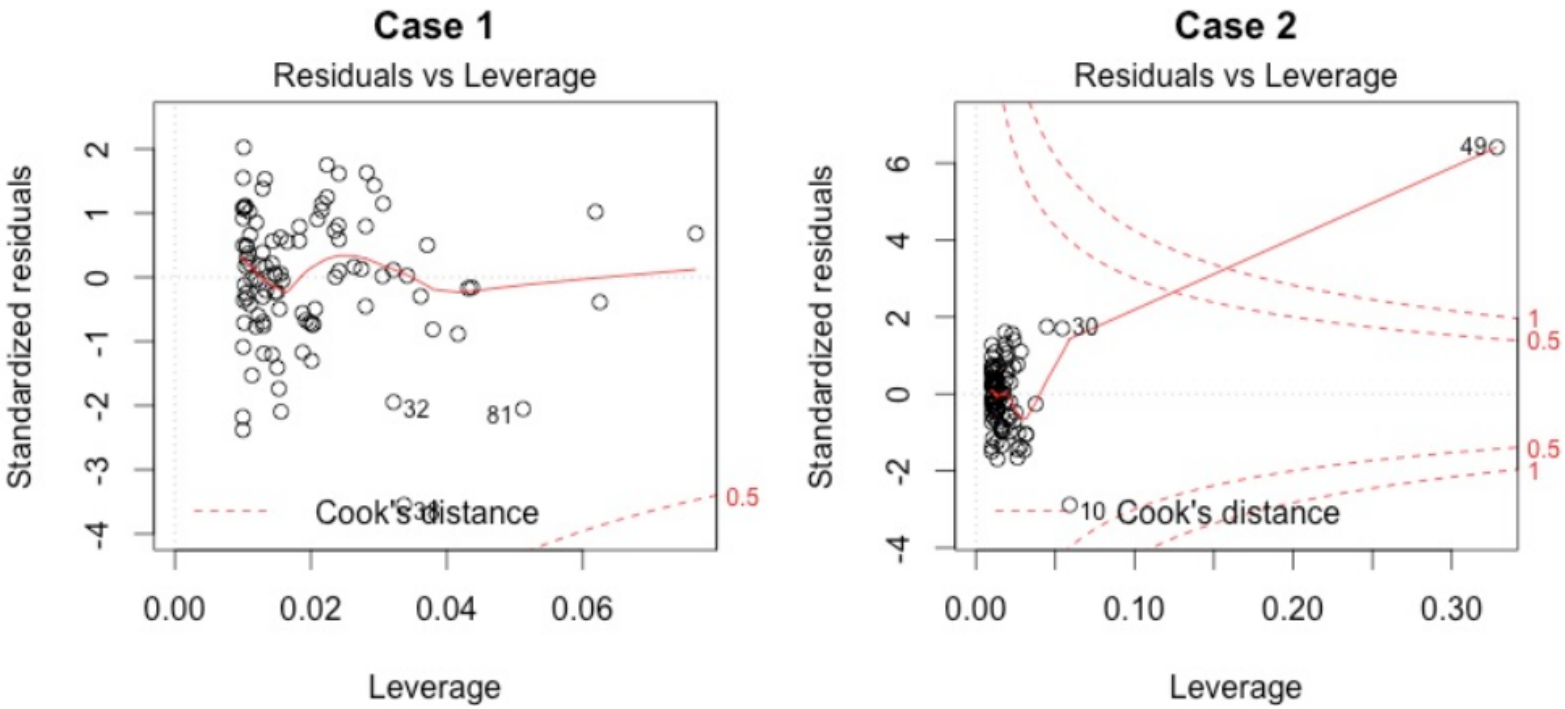
Etudier différents graphiques des résidus pour éviter des erreurs





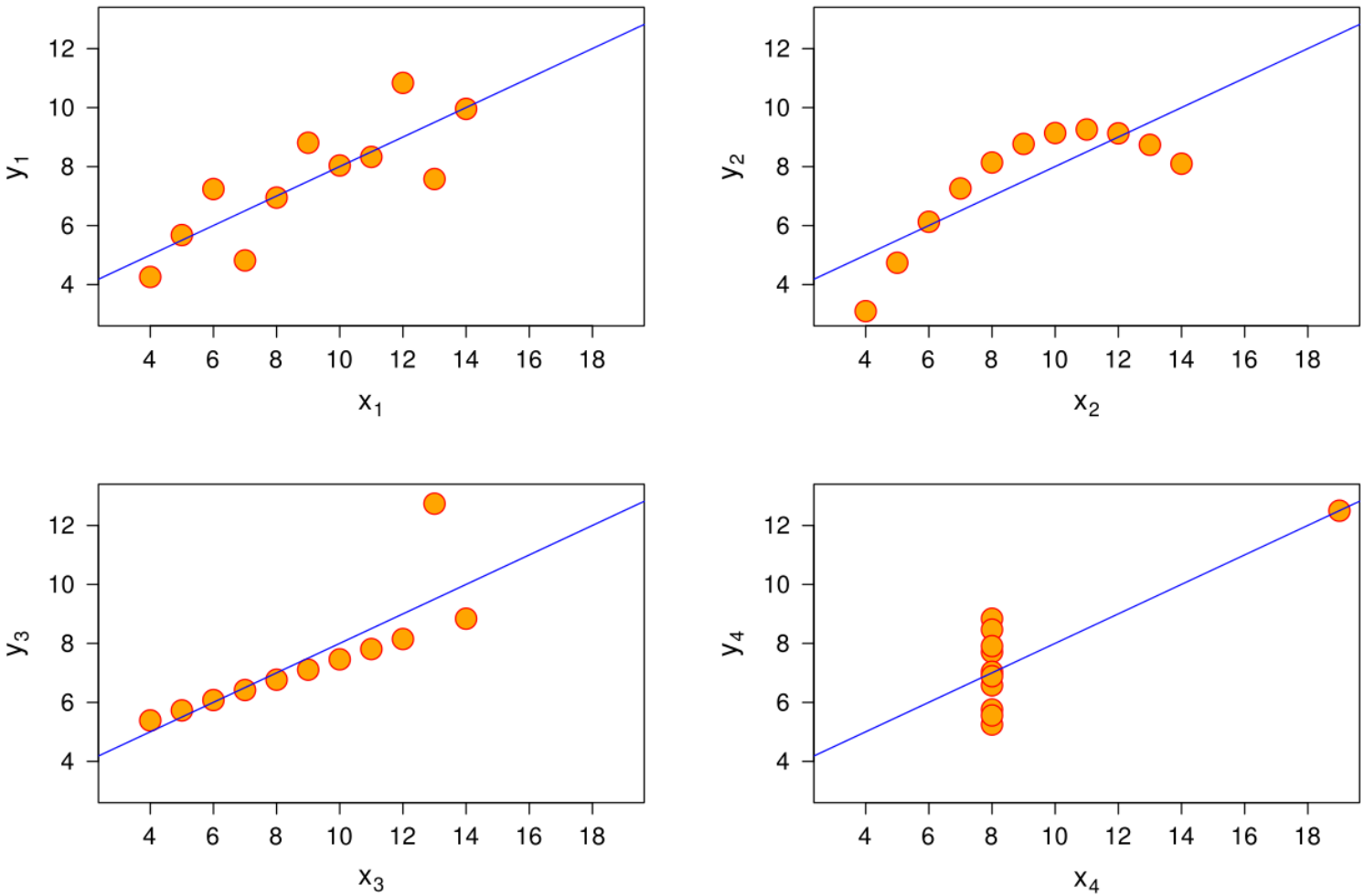
Les pièges à éviter

Etudier différents graphiques des résidus pour éviter des erreurs



Les pièges à éviter

Les règles statistiques à respecter : Le quartet d'Anscombe



1 Introduction

2 Corrélation et régression de variables quantitatives

3 Variables qualitatives

4 Méthodes exploratoires, synthétiques et classification

5 Interprétation des résultats obtenus sous R

6 Introduction à R

Tableau de contingence et valeurs théoriques

Association et indépendance plutôt que corrélation

Lorsque l'on étudie des variables qualitatives, on comprend bien qu'il sera difficile, voire impossible, de produire un nuage de points et par conséquent de calculer des corrélations et des régressions linéaires.

Néanmoins, on peut aussi se dire qu'il faut quand même différencier les variables qualitatives nominales, de celles qui sont ordinales.

On parlera davantage d'association, d'influence, de dépendance ou au contraire d'indépendance dans le cas de variables qualitatives.

Entre deux variables qualitatives, il est par exemple possible de compter les effectifs qui correspondent aux associations (conjonctions) possibles entre les deux variables.

On parle de tableau de contingence. La notion de tableau croisé dynamique, proposée par les tableurs, est une généralisation du tableau de contingence.

Tableau de contingence et valeurs théoriques

Présentation

Les cases du tableau correspondent aux effectifs associés conjointement à une modalité de X et une modalité de Y.

Toutes les modalités de X et de Y y sont représentées.

Il est possible de calculer les valeurs totales du tableau, en ligne et en colonne, qui correspondent aux effectifs marginaux. La somme totale des effectifs correspond à l'effectif global.

A partir des effectifs et des effectifs marginaux, il est possible de calculer des proportions pour chaque ligne (profil en ligne) ou pour chaque colonne (profil en colonne).

La lecture du tableau de contingence sur la base des profils est très instructive, mais en tant que statisticien, il convient de caractériser la force du lien à l'aide d'indicateurs numériques et éventuellement tester si elle est significative.

Tableau de contingence et valeurs théoriques

Présentation

Y × X	x_1	$\cdots$	x_c	$\cdots$	x_C	Total
y_1	n_{11}	$\cdots$	n_{1c}	$\cdots$	n_{1C}	$n_{1.}$
$\vdots$		$\cdots$				$\vdots$
y_l	n_{l1}	$\cdots$	n_{lc}	$\cdots$	n_{lC}	$n_{l.}$
$\vdots$		$\cdots$				$\vdots$
y_L	n_{L1}	$\cdots$	n_{Lc}	$\cdots$	n_{LC}	$n_{L.}$
Total	$n_{.1}$	$\cdots$	$n_{.c}$	$\cdots$	$n_{.C}$	$n = n_{..}$

Tableau de contingence et valeurs théoriques

Exemple

Effectifs observés (Nij)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	130	128	39	14	29	47	21	151	**559**
HONGRIE	144	241	53	28	77	61	91	423	**1118**
POLOGNE	380	612	164	84	222	199	147	881	**2689**
R.D.A.	206	451	119	118	308	142	109	1056	**2509**
ROUMANIE	136	305	244	41	76	114	106	366	**1388**
TCHECO.	185	412	130	63	139	151	177	883	**2140**
YOUGOSL.	126	223	132	58	76	78	69	307	**1069**
Total	**1307**	**2372**	**881**	**406**	**927**	**792**	**720**	**4067**	**11472**

Tableau de contingence et valeurs théoriques

Deux profils possibles : le profil en ligne

Profils en ligne (Nij/Ni.)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	23%	23%	7%	3%	5%	8%	4%	27%	**100%**
HONGRIE	13%	22%	5%	3%	7%	5%	8%	38%	**100%**
POLOGNE	14%	23%	6%	3%	8%	7%	5%	33%	**100%**
R.D.A.	8%	18%	5%	5%	12%	6%	4%	42%	**100%**
ROUMANIE	10%	22%	18%	3%	5%	8%	8%	26%	**100%**
TCHECO.	9%	19%	6%	3%	6%	7%	8%	41%	**100%**
YOUGOSL.	12%	21%	12%	5%	7%	7%	6%	29%	**100%**
Total	**11%**	**21%**	**8%**	**4%**	**8%**	**7%**	**6%**	**35%**	**100%**

Tableau de contingence et valeurs théoriques

Deux profils possibles : le profil en colonne

Profils en colonne (Nij/N.j)									
1963	*ALIM*	*TEXT*	*BOIS*	*EDIT*	*CHIM*	*CONS*	*META*	*EQUIP*	**Total**
BULGARIE	10%	5%	4%	3%	3%	6%	3%	4%	**5%**
HONGRIE	11%	10%	6%	7%	8%	8%	13%	10%	**10%**
POLOGNE	29%	26%	19%	21%	24%	25%	20%	22%	**23%**
R.D.A.	16%	19%	14%	29%	33%	18%	15%	26%	**22%**
ROUMANIE	10%	13%	28%	10%	8%	14%	15%	9%	**12%**
TCHECO.	14%	17%	15%	16%	15%	19%	25%	22%	**19%**
YOUGOSL.	10%	9%	15%	14%	8%	10%	10%	8%	**9%**
Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Tableau de contingence et valeurs théoriques

Valeurs théoriques

Avec un tableau de contingence, on peut donc obtenir la valeur totale des effectifs concernés. $E = 11472$.

On peut aussi obtenir la taille d'une modalité vis-à-vis des autres pour les colonnes. $ALIM = 1307 / 11472 = 0.11$

On peut aussi obtenir la taille d'une modalité vis-à-vis des autres pour les lignes. $BULGARIE = 559 / 11472 = 0.05$

Si l'on multiplie l'ensemble de ces valeurs, on obtient une valeur théorique, qui correspond à ce que l'on pourrait obtenir si les deux variables étaient indépendantes. $11472 \times 0.11 \times 0.05 = 63$

Cette valeur correspond à ce que l'on pourrait s'attendre à obtenir si la situation était « simple » : sans dépendance, sans sur-représentation, sans sous-représentation, sans spécificité locale...

Le rapport entre la valeur réelle (130) et la valeur théorique (63), c'est ce que mesure le quotient de localisation.

Tableau de contingence et valeurs théoriques

Valeurs théoriques

Valeur Théorique	ALIM	TEXT	BOIS	EDIT	CHIM	CONS	META	EQUIP
BULGARIE	63,69	115,58	42,93	19,78	45,17	38,59	35,08	198,17
HONGRIE	127,37	231,16	85,86	39,57	90,34	77,18	70,17	396,35
POLOGNE	306,36	555,99	206,50	95,17	217,29	185,64	168,77	953,29
R.D.A.	285,85	518,77	192,68	88,79	202,74	173,22	157,47	889,48
ROUMANIE	158,13	286,99	106,59	49,12	112,16	95,82	87,11	492,07
TCHECO.	243,81	442,48	164,34	75,74	172,92	147,74	134,31	758,66
YOUGOSL.	121,79	221,03	82,09	37,83	86,38	73,80	67,09	378,98

Test du Chi-2

Principes

L'idée du chi-2 (χ^2) de Pearson est de comparer les effectifs réellement observés (o_k) avec les effectifs théoriques (e_k) si les variables X et Y étaient indépendantes.

Pour cela, cette technique s'appuie sur une mesure, appelée mesure du χ^2. La statistique du χ^2 quantifie l'écart (la distance) entre tous les effectifs observés et tous les effectifs théoriques.

$$\chi^2 = \sum_{k}^{K} \frac{(o_k - e_k)^2}{e_k}$$

Dans notre cas, la première valeur de ce calcul du χ^2 est :

$$(130 - 63,69)^2/63,69 = 69$$

Test du Chi-2

Calcul

Chi-2	ALIM	TEXT	BOIS	EDIT	CHIM	CONS	META	EQUIP
BULGARIE	69,05	1,33	0,36	1,69	5,79	1,83	5,65	11,23
HONGRIE	2,17	0,42	12,57	3,38	1,97	3,39	6,19	1,79
POLOGNE	17,70	5,64	8,75	1,31	0,10	0,96	2,81	5,48
R.D.A.	22,31	8,85	28,18	9,61	54,65	5,63	14,92	31,17
ROUMANIE	3,10	1,13	177,13	1,34	11,66	3,45	4,09	32,30
TCHECO.	14,19	2,10	7,18	2,14	6,66	0,07	13,57	20,38
YOUGOSL.	0,15	0,02	30,34	10,75	1,25	0,24	0,05	13,67
							Total	**703,82**

Cette valeur totale peut alors faire l'objet d'un test d'indépendance en s'appuyant sur une table du χ^2. Il faut pour cela définir un niveau de risque. Pour déterminer le nombre de degrés de liberté, il faut effectuer le calcul suivant où Nc est le nombre de colonnes et Nl le nombre de lignes :

$$DL = (Nc - 1) \times (Nl - 1)$$

Si la valeur du χ^2 est supérieure à celle du tableau alors les deux variables sont liées. Les logiciels fournissent souvent la p-value.

Test du Chi-2

Table du χ^2

dl	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
10	2.15	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.75
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.21	28.30
13	3.56	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.69	26.12	29.14	31.31
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.15
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.83	28.41	31.41	34.17	37.56	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	20.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	21.34	26.04	30.81	33.93	36.78	40.29	42.80
23	9.26	10.19	11.69	13.09	14.85	18.14	22.34	27.14	32.01	35.17	38.08	41.64	44.18
24	9.88	10.86	12.40	13.85	15.66	19.04	23.34	28.24	33.20	36.42	39.37	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	24.34	29.34	34.38	37.65	40.65	44.32	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	25.34	30.43	35.56	38.89	41.92	45.64	48.29
27	11.80	12.88	14.57	16.15	18.11	21.75	26.34	31.53	36.74	40.11	43.20	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	27.34	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	28.34	33.71	39.09	42.56	45.72	49.59	52.34
30	13.78	14.95	16.79	18.49	20.60	24.48	29.34	34.80	40.26	43.77	46.98	50.89	53.67
40	20.67	22.14	24.42	26.51	29.06	33.67	39.34	45.61	51.80	55.75	59.34	63.71	66.80
50	27.96	29.68	32.35	34.76	37.69	42.95	49.34	56.33	63.16	67.50	71.42	76.17	79.52
60	35.50	37.46	40.47	43.19	46.46	52.30	59.34	66.98	74.39	79.08	83.30	88.40	91.98
70	43.25	45.42	48.75	51.74	55.33	61.70	69.34	77.57	85.52	90.53	95.03	100.44	104.24
80	51.14	53.52	57.15	60.39	64.28	71.15	79.34	88.13	96.57	101.88	106.63	112.34	116.35
90	59.17	61.74	65.64	69.13	73.29	80.63	89.33	98.65	107.56	113.14	118.14	124.13	128.32
100	67.30	70.05	74.22	77.93	82.36	90.14	99.33	109.14	118.49	124.34	129.56	135.82	140.19

Test du Chi-2

Conclusion

Dans notre exemple, le nombre de degrés de liberté est de : (8-1) $\times$ (7-1) = 42. D'après la table du χ^2, pour un risque de 5 % et un nombre de degrés de liberté de 42, la valeur de référence est comprise entre 55,75 et 67,50.

La valeur du χ^2 est donc très largement supérieure à la valeur de référence. La localisation et la production sont liées.

Attention

Facile à utiliser le test du χ^2 doit en théorie remplir certaines conditions d'application : un effectif global suffisant (>20), peu d'effectifs faibles (80 % des cases $>$ 5).

Lorsque les effectifs sont très élevés, le test du χ^2 aboutit presque systématiquement au rejet de l'hypothèse d'indépendance. Un petit écart, aussi infime soit-il, se répercute fortement sur la statistique.

Test du Chi-2

Alternatives

Si le test du χ^2 est très répandu, il existe néanmoins des alternatives.

Le test du V de Cramer, qui s'appuie sur la métrique du χ^2, permet d'obtenir une valeur de l'intensité de la liaison.

En épidémiologie, il est classique de calculer des ODDS ratios (rapports de cotes), une des deux variables qualitatives doit être de type binaire (malade/sain).

Lorsque les deux variables sont binaires, il est pertinent de passer par un test de corrélation (0 pour absence non, 1 pour présence oui).

Enfin, le χ^2 peut constituer une alternative à la corrélation, ce n'est pas un appauvrissement surtout lorsqu'une variable quantitative a un comportement discret (années d'études supérieures, durée de prêts...).

Anova

Présentation

En statistique, l'analyse de la variance (ANOVA : analysis of variance) est un ensemble de modèles statistiques utilisés pour vérifier si les moyennes de différents groupes sont égales.

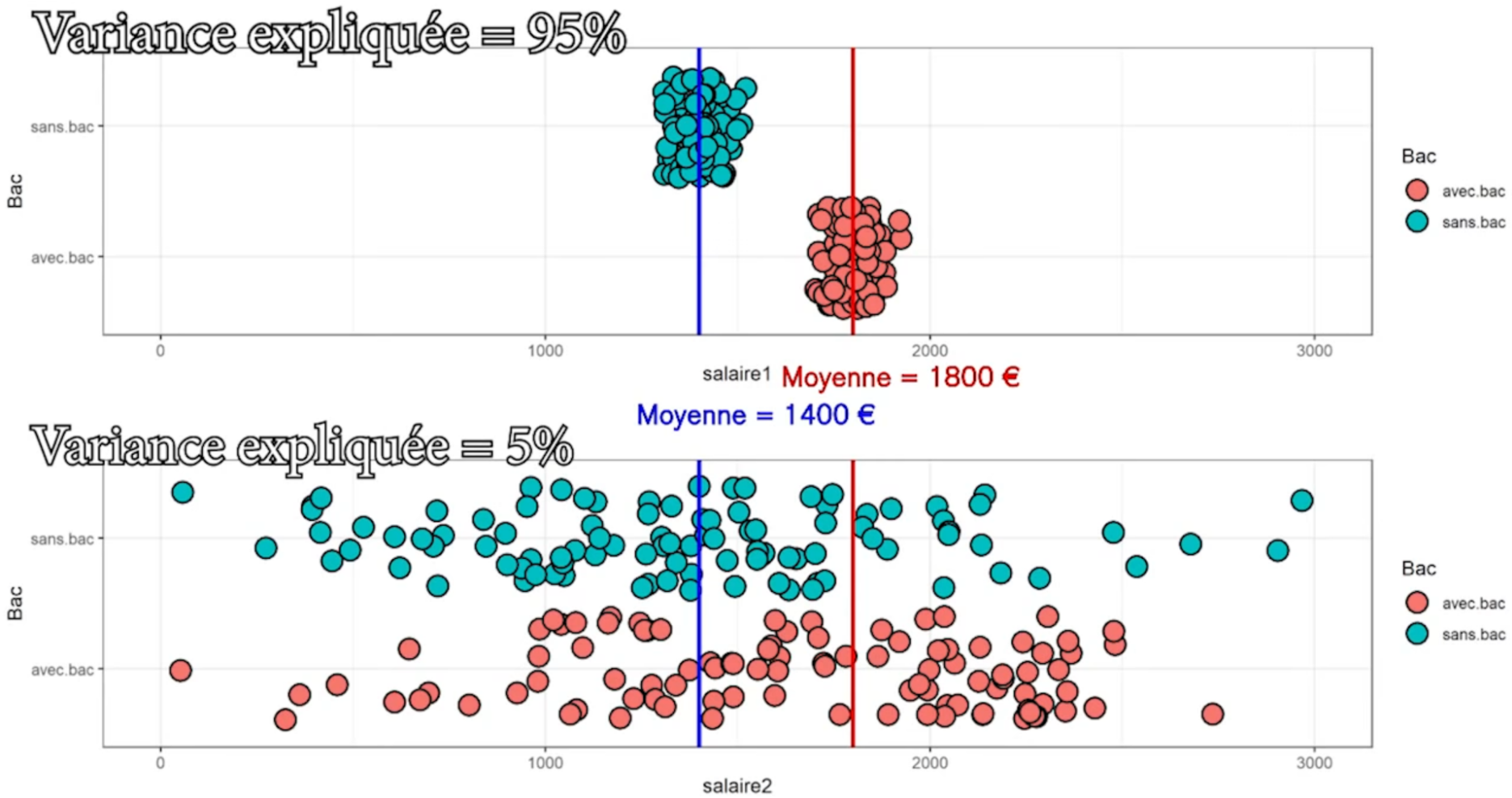
Cette analyse est appelée « analyse de variance » car sa procédure s'appuie sur les variances pour déterminer si les moyennes sont différentes.

Ce test s'applique lorsque l'on mesure une ou plusieurs variables explicatives catégorielle (appelées alors facteurs de variabilité, leurs différentes modalités étant parfois appelées « niveaux ») qui peuvent avoir de l'influence sur une variable quantitative continue.

On parle d'analyse à un facteur lorsque l'analyse porte sur un modèle décrit par un seul facteur de variabilité, d'analyse à deux facteurs ou d'analyse multifactorielle sinon (MANOVA).

Anova

Exemple



Anova

Détails statistiques

Les écarts à la moyenne qui interviennent dans le calcul de la variance peuvent s'écrire de la manière suivante :

$$X_{ij} - \bar{X} = (\bar{X}_j - \bar{X}) + (X_{ij} - \bar{X}_j)$$

Avec ce petit jeu d'écriture, qui introduit la moyenne $\bar{X}_j$ d'un facteur dans la formule, on écrit que l'écart à la moyenne globale est égal à l'écart entre les groupes plus l'écart à l'intérieur des groupes. On obtient alors la formule de variance suivante :

$$\sum_{i}^{nj}\sum_{j}^{p}(X_{ij} - \bar{X})^2 = \sum_{i}^{nj}\sum_{j}^{p}(\bar{X}_j - \bar{X})^2 + \sum_{i}^{nj}\sum_{j}^{p}(X_{ij} - \bar{X}_j)^2$$

Le premier terme calcule la variance globale (correspondant au SCT), le second terme calcule la variance expliquée par les moyennes des groupes (SCE), le troisième terme calcule la variance au sein des groupes la variance résiduelle (SCR).

Anova

Détails statistiques

Très souvent les logiciels donnent les carrés moyens et non les sommes des carrés moyens :

$$\text{CMT} = \text{SCT} \ / \ (\text{n-1}) \, ; \ \text{CME} = \text{SCE} \ / \ (\text{p - 1}) \, ; \ \text{CMR} = \text{SCR} \ / \ (\text{n - p})$$

A partir de ces valeurs, un test de Fisher peut être effectué pour déterminer la significativité de ces écarts pour un risque donné :

$$F = CME/CMR \text{ valeur à comparer à } Fischer(p-1, n-p)$$

Le rapport SCE/SCT peut être interprété comme un R^2.

Ces calculs peuvent être généralisés à plusieurs facteurs. On introduira pour cela les variations des interactions entre différents facteurs.

L'Anova peut permettre d'effectuer des tests de Fisher pour comparer des modèles emboîtés (les variables du plus petit modèle sont contenues dans le plus grand modèle).

Ancova

Présentation

Le défaut de l'Anova est qu'elle ne s'applique qu'à des variables explicatives qualitatives. Le défaut des régressions présentées est qu'elles ne s'appliquent qu'à des variables explicatives quantitatives.

Or, très souvent les deux types de variables se mélangent. Un exemple est le rapport prix/superficie des logements. Quels que soient les territoires (variables qualitatives) le prix à tendance à augmenter de manière linéaire avec la superficie (le fameux prix au mètre carré).

Si vous prenez tous les territoires de manière indifférente dans un modèle de régression, vous pouvez obtenir n'importe quoi (car le prix au mètre carré n'est pas le même partout et les contraintes sur les logements non plus).

De même, si vous faites une Anova sur les relations territoire/prix ou territoire/superficie, il faut le faire en tenant compte de la relation prix/superficie au sein de chaque territoire au risque de sous-évalué les deux liens étudiés. L'Ancova propose une solution pour résoudre ce type de problème.

Ancova

Présentation

L'analyse de la covariance (ANCOVA) est une méthode statistique visant à tester, par un modèle linéaire général, l'effet sur une variable dépendante continue d'une ou plusieurs variables indépendantes catégorielles, indépendamment de l'effet d'autres facteurs quantitatifs continus (de covariables).

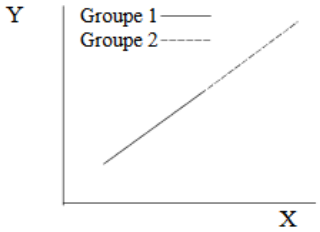
En d'autres termes, l'Ancova est une combinaison entre une Anova et une régression linéire, de telle sorte que l'Ancova permette de tester si certains facteurs ont un effet sur la variable à expliquer après avoir enlevé la variance due aux covariables.

L'Ancova permet donc en quelque sorte de comparer des moyennes ajustées de deux ou plusieurs groupes indépendants (toute chose égale par ailleurs).

Ancova

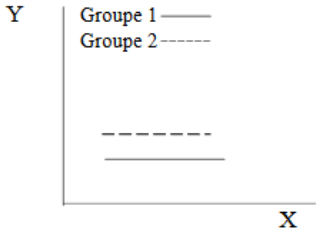
Différentes configurations

Y: variable dépendante; X: co-facteur (Prédicteur Continu); G: variable indépendante (Prédicteur Catégoriel; discret). On teste les effets de X, G et X*G (interaction) sur la variable Y



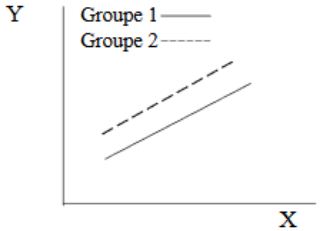
Cas 1 : X est significatif, G et X*G ne le sont pas.

Y change en changeant X, alors X a un effet significatif sur Y. Par contre, les deux points d'intersection et les deux pentes sont les mêmes.



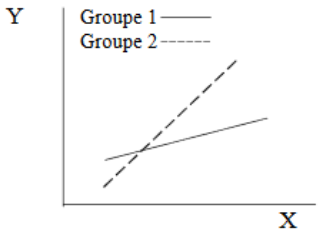
Cas 2 : G est significatif, X et X*G ne le sont pas.

Y ne change pas en changeant X, alors X n'a pas d'effet sur Y. Les points d'intersection des deux groupes sont différents, alors G a un effet significatif sur Y. Par contre, les deux pentes sont égales (zéro) donc G*X n'a pas d'effet sur Y.



Cas 3 : G et X sont significatifs, X*G ne l'est pas.

Y change en changeant X, alors X affecte Y. Les points d'intersection des deux groupes sont différents, alors G affecte Y également. Par contre, les deux pentes sont égales (les lignes sont parallèles) donc l'effet de Y sur X ne varie pas en fonction de la valeur de G (c'est-à-dire, dépendant du groupe). Alors X*G n'est pas significatif.



Cas 4 : G, X et X*G sont significatifs.

Y change en changeant X, alors X affecte Y. Les points d'intersection des deux groupes sont différents, alors G affecte Y également. En plus, les deux pentes sont différentes (les lignes ne sont pas parallèles) donc l'effet de Y sur X dépend de la valeur de G (c'est-à-dire, dépend du groupe). Alors X*G est significatif.

Ancova

Différentes hypothèses

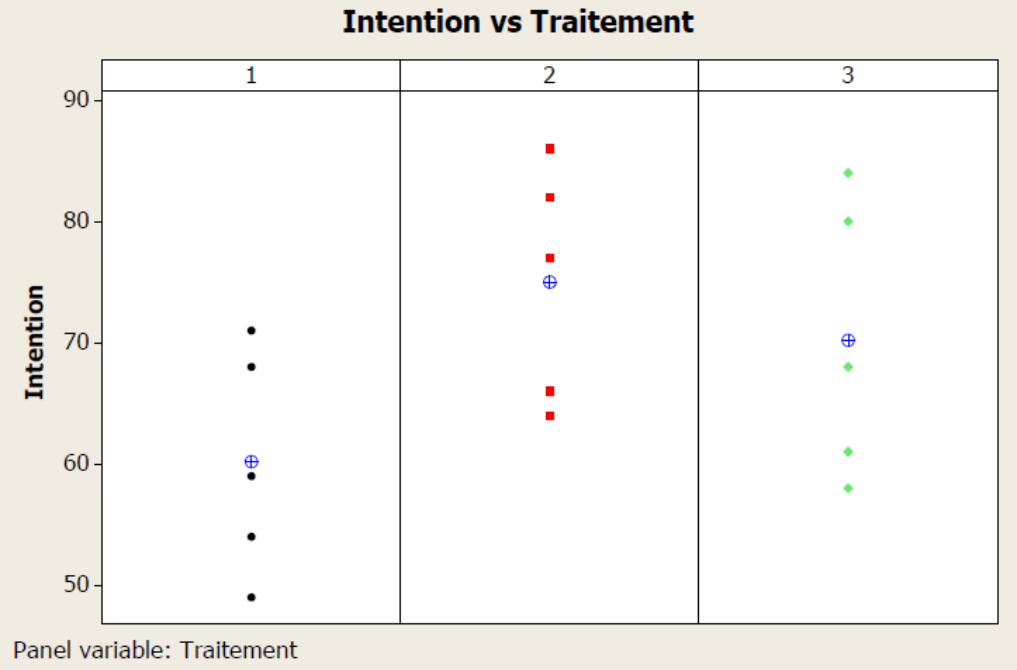
L'Ancova fait plusieurs hypothèses au sujet des données :

- linéarité entre la covariable et la variable-réponse à chaque niveau de la variable de groupement. Diagramme de dispersion groupé de la covariable et de la variable-réponse.
- homogénéité des pentes de régression. Les pentes des droites de régression devraient être les mêmes pour chaque groupe. Cette hypothèse évalue qu'il n'y a pas d'interaction entre le résultat et la covariable.
- la variable-réponse doit être approximativement distribuée normalement. Test de normalité Shapiro-Wilk sur les résidus du modèle.
- homoscedasticité ou homogénéité de la variance des résidus pour tous les groupes.
- aucune valeur aberrante dans les groupes.

Ancova

Exemple

Prenons le cas de l'évaluation de trois publicités sur l'intention d'achat d'un produit.

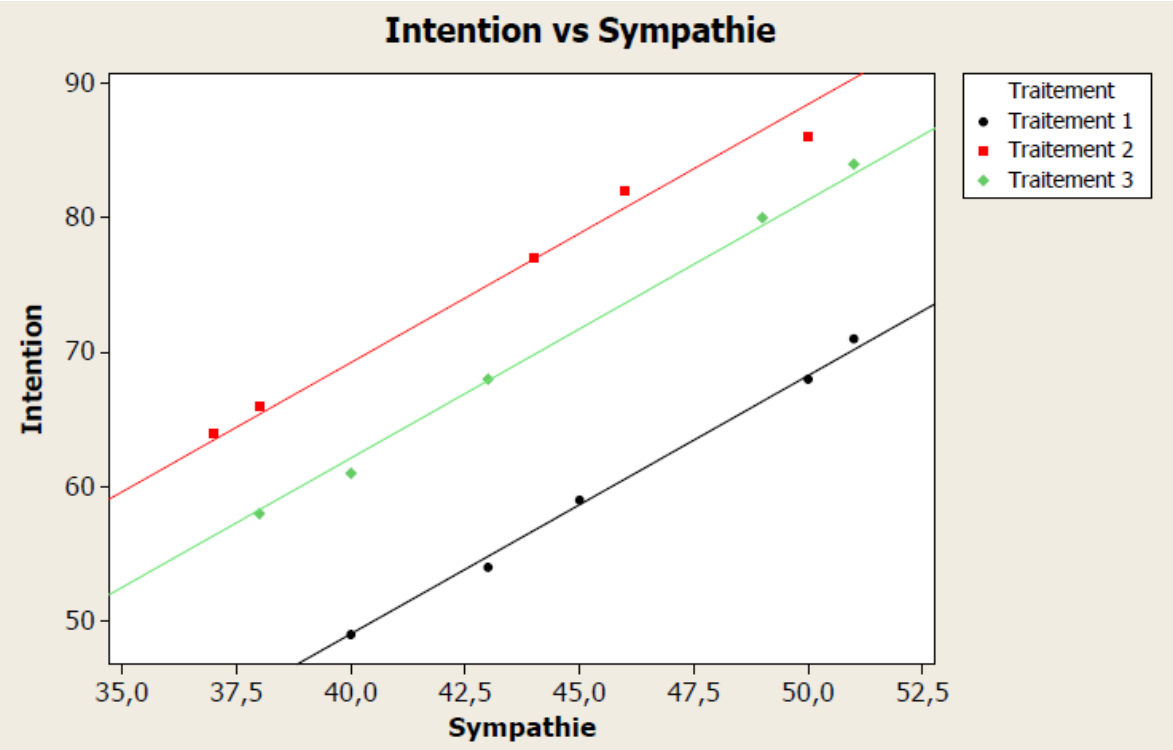


D'après ce graphique, il semble que la publicité 2 soit en moyenne la plus efficace en termes d'intention d'achat. Une Anova nous permettrait de déterminer si cette différence est significative pour conclure cela.

Ancova

Exemple

Néanmoins, cela serait trop simple, car les personnes avaient déjà une certaine intention d'achat avant la publicité (une certaine sympathie).



La question est alors la suivante, quelle est la publicité réellement la plus efficace ?

Ancova

Réalisation

L'Ancova consiste à trouver une relation de type linéaire avec des variables explicatives qualitatives et quantitatives.

Il faudra réaliser une Anova (un test F) sur le modèle Ancova de façon à vérifier que toutes les variables sont bien explicatives.

Il faudra de surcroit s'intéresser à l'existence (et à la prise en compte) des interactions entres les variables explicatives toujours à l'aide d'une Anova sur les modèles avec ou sans interactions, cela permettra de savoir si les coefficients directeurs des régressions peuvent être considérés comme identiques.

On pourra alors calculer les moyennes ajustées qui seront égales à :

$$\bar{Y}_j = constante \pm coefquali + \bar{X} * coefquanti$$

Régression logistique

Présentation

L'Ancova permet de comprendre qu'il est pertinent de travailler avec des modèles de régression lorsque l'on dispose de variables explicatives qualitatives.

Néanmoins, l'Ancova ne permet pas de résoudre le cas où c'est la variable à expliquer qui est qualitative. Dans ce cadre, on utilisera les modèles de régression logistique (modèle logit) ou de régression multinomiale.

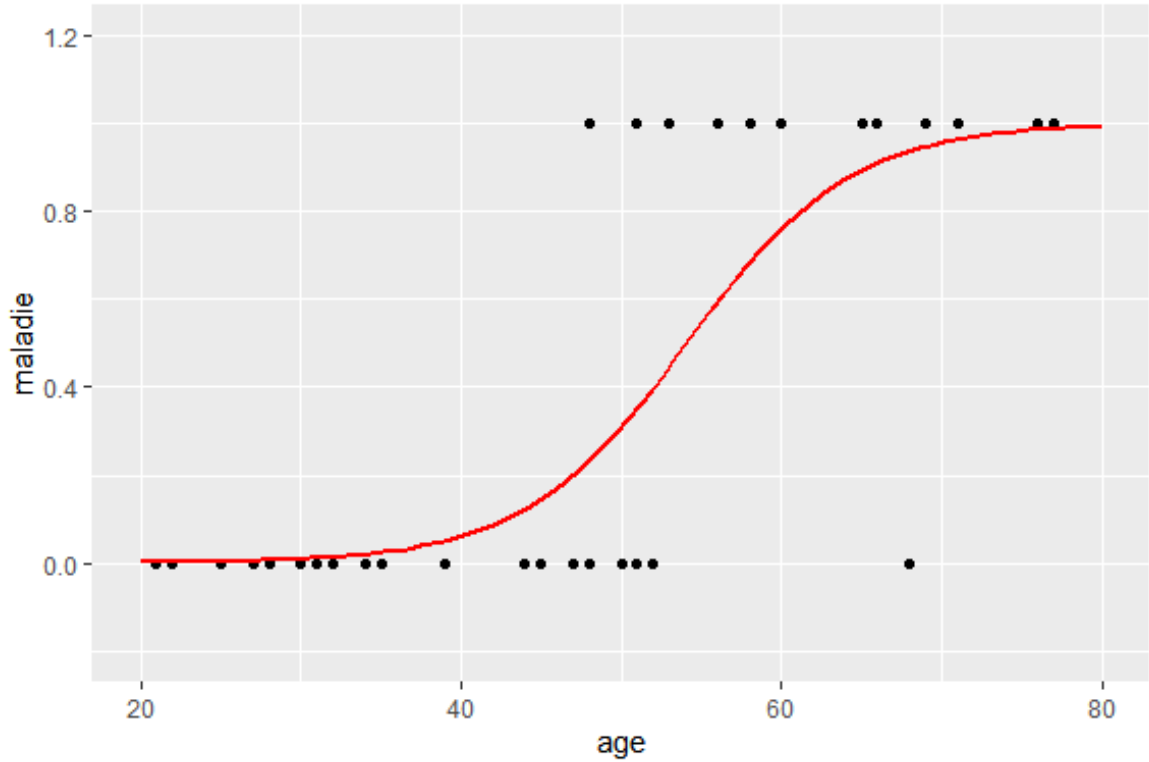
La régression logistique s'applique au cas où la variable à expliquer est de type binaire.

$$P(x) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}$$

Les régressions multinomiales sont une généralisation des régressions logistiques. Ces régressions sont des exemples de modèles linéaires généralisés.

Régression logistique

Exemple



Dans cet exemple, la valeur β_1 est égale à 0,19 et est significativement supérieure à zéro. L'âge apparait être un facteur de risque.

L'ODDS ratio, le coefficient directeur, la taille de l'effet se calcule en prenant l'exponentiel de cette valeur : 1,209.

1 Introduction

2 Corrélation et régression de variables quantitatives

3 Variables qualitatives

4 Méthodes exploratoires, synthétiques et classification

5 Interprétation des résultats obtenus sous R

6 Introduction à R

ACP

Présentation

L'Analyse en Composantes Principales (ACP) est une méthode de la famille de la statistique multivariée qui consiste à transformer des variables liées entre elles (corrélées) en nouvelles variables décorrélées les unes des autres (indépendantes).

Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Ces composantes cherchent alors à restituer aux mieux les variations du jeu de données. On va alors utiliser ces méthodes pour représenter l'information avec moins de composantes principales que de variables.

Cette méthode permet en quelque sorte de réduire le nombre de variables en rendant l'information moins redondante. On va pour cela accepter de perdre un peu d'information, car cela va permettre de simplifier l'interprétation.

Généralement, on choisira une représentation comportant deux composantes principales.

Cette méthode s'applique uniquement à des variables quantitatives.

ACP

Présentation

L'ACP se rattache à la famille des analyses factorielles qui regroupe différentes méthodes d'analyses de grands tableaux rectangulaires de données, visant toutes à identifier et à hiérarchiser des facteurs corrélés aux données.

L'ACP s'applique très bien à des tableaux d'information géographique, puisqu'elle s'appuie sur des tableaux avec en lignes des individus et en colonnes des variables.

Dans un tableau d'information géographique, les individus sont des entités géographiques. C'est pourquoi cette méthode est très utilisée en géographie.

L'objectif est alors de simplifier l'information pour permettre par exemple d'identifier plus facilement des ressemblances entre les entités géographiques.

Néanmoins, l'objectif sans doute premier de l'ACP, c'est d'analyser les liaisons entre les variables et d'identifier les redondances (les corrélations).

ACP

Exemple

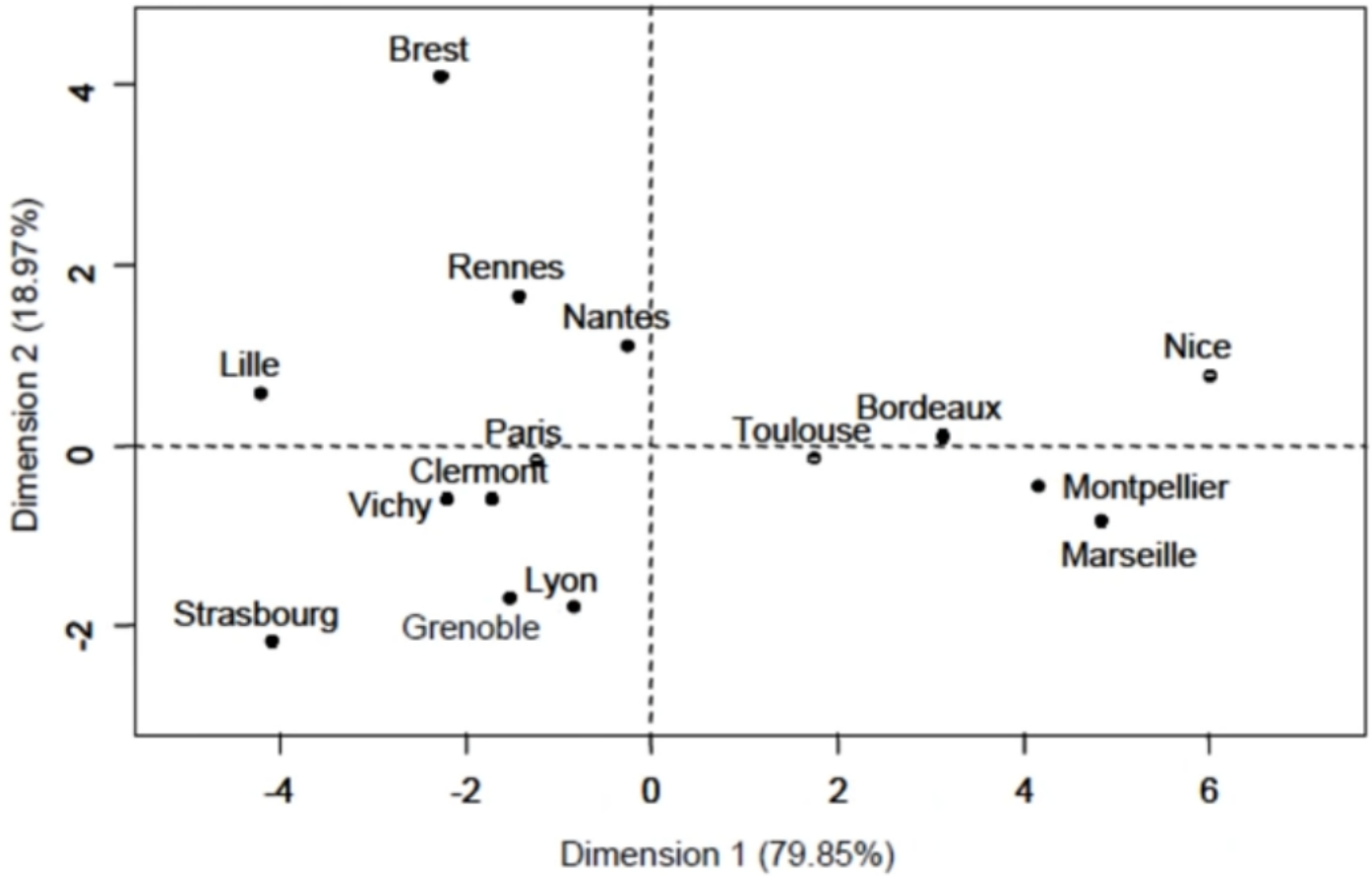
	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

A partir de ces données, on peut se demander quelles villes ont des profils de températures similaires ou au contraire opposées.

ACP

Exemple

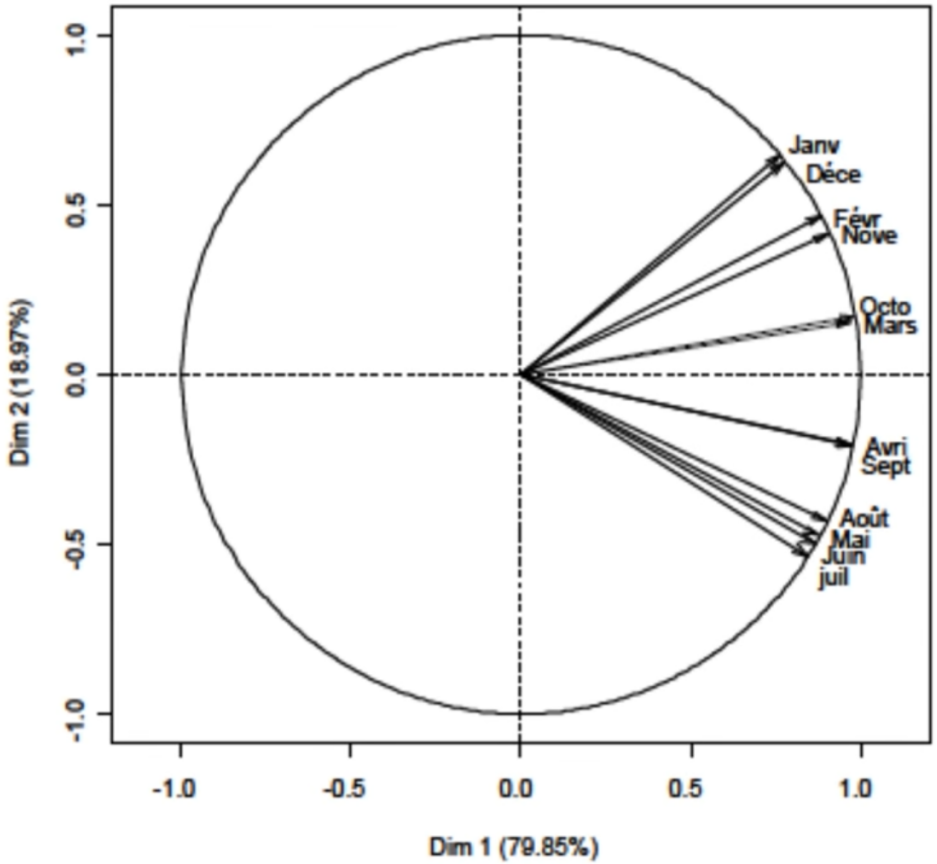
En choisissant de représenter les individus sur un graphique fondé sur les axes des deux premières composantes principales de l'ACP, on peut en partie répondre à cette question.



ACP

Exemple

On peut faire la même chose pour les variables, c'est le cercle des corrélations. L'ACP peut alors permettre de ne pas étudier deux à deux toutes les corrélations pour identifier des variables redondantes.



AFC, ACM et AFM

Présentation

L'AFC (Analyse Factorielle des Correspondances) se différencie de l'ACP en ce sens qu'elle s'applique uniquement à des tableaux de contingence appelés tableaux de correspondance.

Ainsi, l'AFC peut être présentée comme une solution pour appliquer une analyse factorielle à des variables qualitatives.

Dans ce cadre, le concept de similarité entre les lignes et les colonnes est différent, car la similarité entre deux lignes ou deux colonnes est complètement symétrique. Deux lignes sont proches l'une de l'autre si elles s'associent aux colonnes de la même façon.

Une fois encore on pourra utiliser cette analyse sur des données géographiques.

AFC, ACM et AFM

Présentation

Les liens entre AFC et méthodes du χ^2 sont forts, mais l'AFC ne traite pas la question de la significativité de la liaison et s'intéresse uniquement à la nature de cette liaison.

L'ACM (Analyse des Correspondances Multiples) permettra d'étudier plusieurs variables qualitatives.

Dans l'ACM, on retrouve en lignes des individus et en colonnes les variables qualitatives. Ces problématiques sont alors presque les mêmes que celles de l'ACP.

Attention, l'ACM passera par la production d'un tableau disjonctif complet, qui s'applique au sens strict du terme à des individus. Ainsi son application à des tableaux d'information géographique est complexe.

L'AFM (Analyse Factorielle Multiple) est une généralisation des méthodes factorielles et pourra s'appliquer à des variables qualitatives et quantitatives.

AFC, ACM et AFM

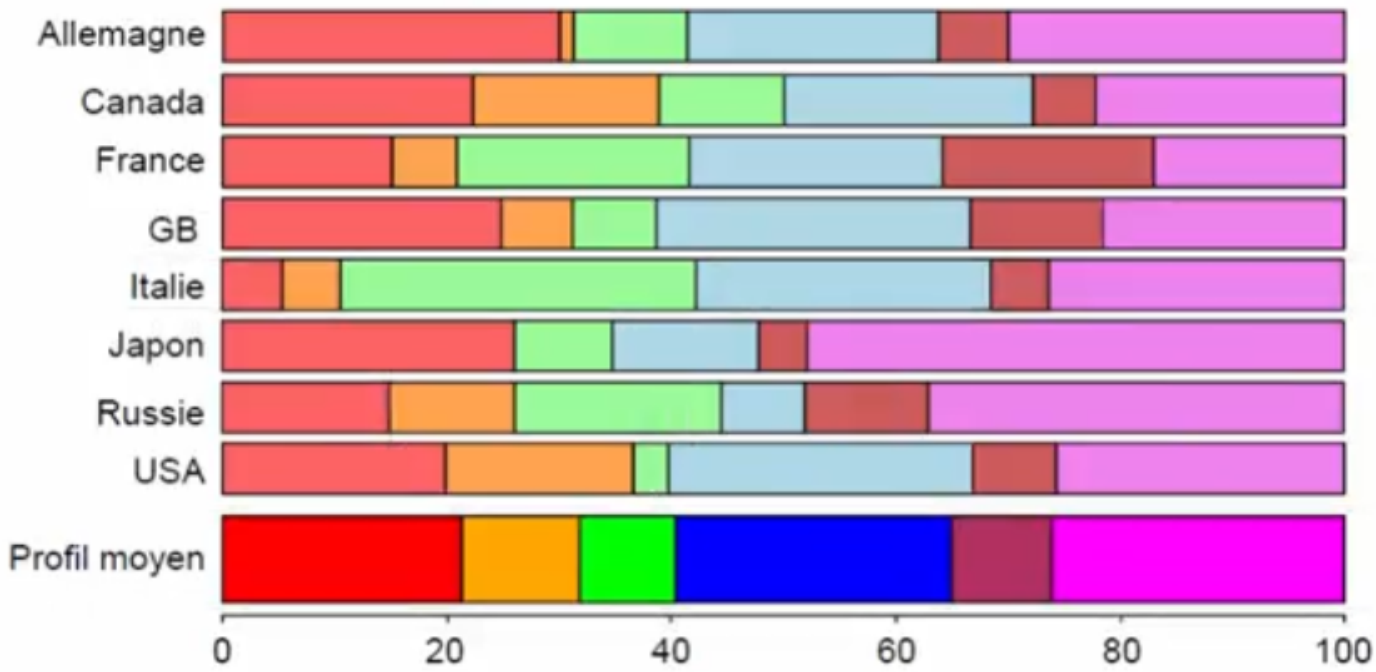
Exemple AFC

	Chimie	Economie	Littérature	Medecine	Paix	Physique	Somme
Allemagne	24	1	8	18	5	24	80
Canada	4	3	2	4	1	4	18
France	8	3	11	12	10	9	53
GB	23	6	7	26	11	20	93
Italie	1	1	6	5	1	5	19
Japon	6	0	2	3	1	11	23
Russie	4	3	5	2	3	10	27
USA	51	43	8	70	19	66	257
Somme	121	60	49	140	51	149	570

AFC, ACM et AFM

Exemple AFC

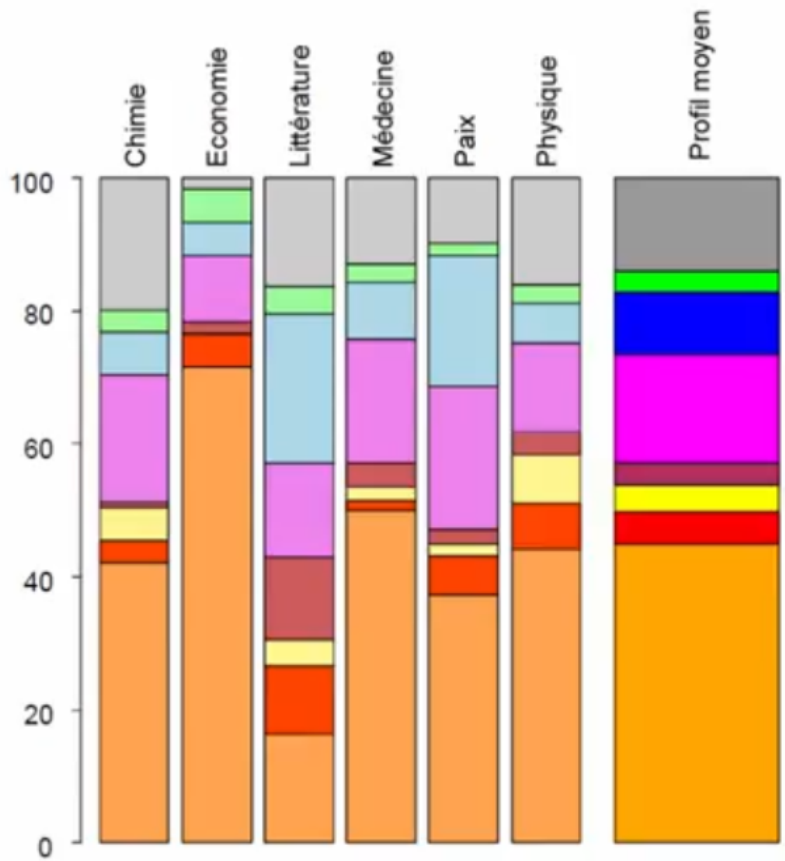
	Chimie	Eco	Lit.	Médecine	Paix	Physique	Somme
Allemagne	30.0	1.2	10.0	22.5	6.2	30.0	100
Canada	22.2	16.7	11.1	22.2	5.6	22.2	100
France	15.1	5.7	20.8	22.6	18.9	17.0	100
GB	24.7	6.5	7.5	28.0	11.8	21.5	100
Italie	5.3	5.3	31.6	26.3	5.3	26.3	100
Japon	26.1	0.0	8.7	13.0	4.3	47.8	100
Russie	14.8	11.1	18.5	7.4	11.1	37.0	100
USA	19.8	16.7	3.1	27.2	7.4	25.7	100
Profil moyen	21.2	10.5	8.6	24.6	8.9	26.1	100



AFC, ACM et AFM

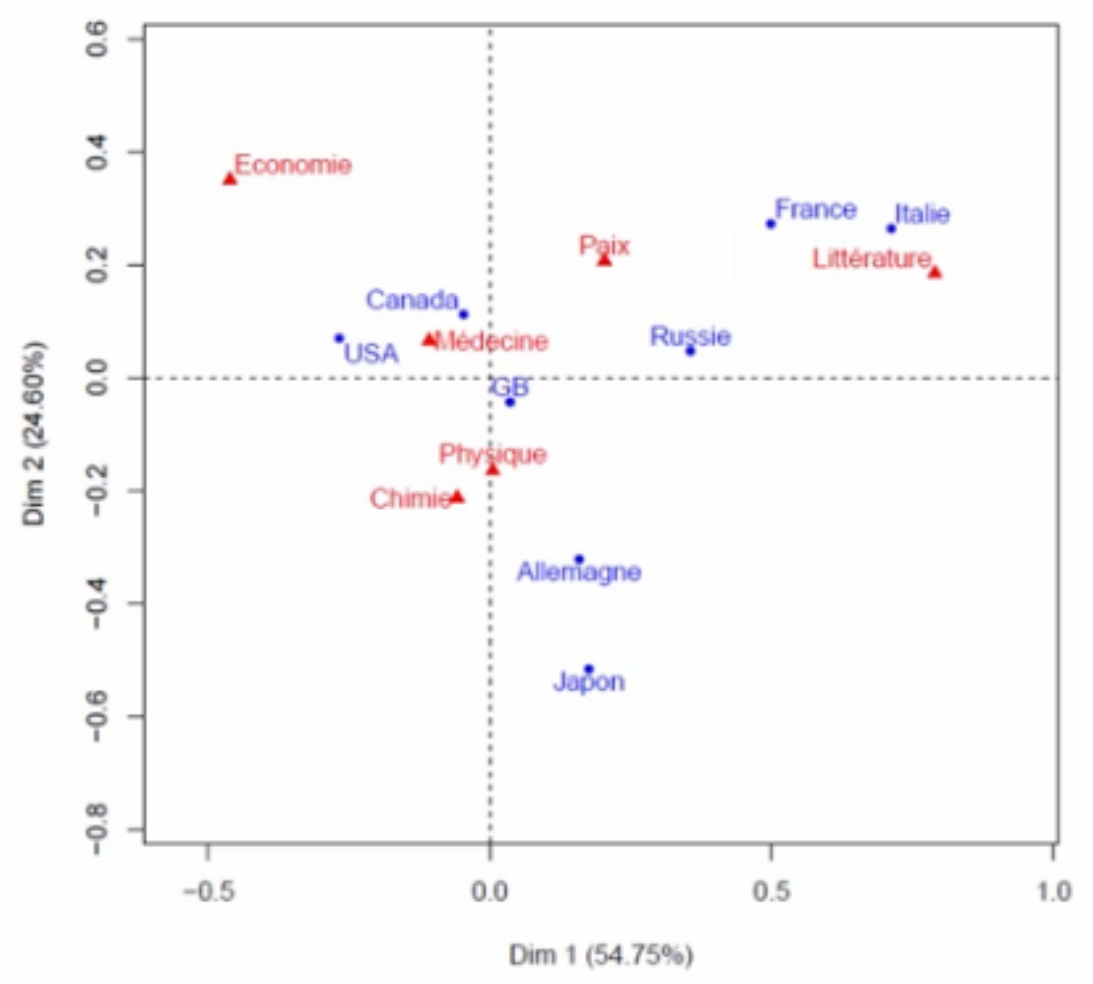
Exemple AFC

	Chimie	Eco	Lit	Méd	Paix	Phys	Profil moyen
Allemagne	19.8	1.7	16.3	12.9	9.8	16.1	14.0
Canada	3.3	5.0	4.1	2.9	2.0	2.7	3.2
France	6.6	5.0	22.4	8.6	19.6	6.0	9.3
GB	19.0	10.0	14.3	18.6	21.6	13.4	16.3
Italie	0.8	1.7	12.2	3.6	2.0	3.4	3.3
Japon	5.0	0.0	4.1	2.1	2.0	7.4	4.0
Russie	3.3	5.0	10.2	1.4	5.9	6.7	4.7
USA	42.1	71.7	16.3	50.0	37.3	44.3	45.1
Somme	100	100	100	100	100	100	100



AFC, ACM et AFM

Exemple AFC



CAH

Présentation

La Classification Ascendante Hiérarchique (CAH) est une méthode de classification itérative dont le principe est simple, l'objectif étant de regrouper des objets au sein de classes (les objets qui se ressemblent dans une même classe, les objets dissemblables dans des classes différentes) :

- On commence par calculer la dissimilarité entre les objets (individus).
- Puis on regroupe les deux objets les plus similaires, créant ainsi une classe comprenant ces deux objets.
- On calcule ensuite la dissimilarité entre cette classe et les autres objets en utilisant un critère d'agrégation. Puis on regroupe les deux objets ou classes d'objets les plus similaires.
- On continue ainsi jusqu'à ce que tous les objets soient regroupés.

Attention, cette méthode est sensible à la redondance des variables étudiées. Par conséquent, il peut être très pertinent de l'utiliser sur les coordonnées d'une analyse factorielle.

CAH

Calcul de dissimilarité

	Xi (km)	Yi (km)
Paris	600	2428
Marseille	846	1815
Saint-Etienne	760	2050
Bordeaux	369	1986
Reims	723	2474
Lyon	794	2087

$$Distance(euclidienne) = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

$$Dist_{(Paris-Marseille)} = \sqrt{(600 - 846)^2 + (2428 - 1815)^2} = 660$$

CAH

Calcul de dissimilarité

Distance euclidienne : $\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$

$De_{(P-M)} = \sqrt{(600 - 846)^2 + (2428 - 1815)^2} = 660$

Distance de Manhattan : $\mid X_1 - X_2 \mid + \mid Y_1 - Y_2 \mid$

$Dm_{(P-M)} = \mid 600 - 846 \mid + \mid 2428 - 1815 \mid = 246 + 613 = 859$

Distance de Tchebychev : $Max[(X_1 - X_2); (Y_1 - Y_2)]$

$Dt_{(P-M)} = Max[(600 - 846); (2428 - 1815)] = Max[246; 613] = 613$

CAH

Calcul de dissimilarité

	Variable 1	Variable 2	Variable 3	Variable 4
Objet 1	5	2	6	4
Objet 2	2	5	2	4

$$Dist_{(Paris-Marseille)} = \sqrt{(5-2)^2 + (2-5)^2 + (6-2)^2 + (4-4)^2}$$

CAH

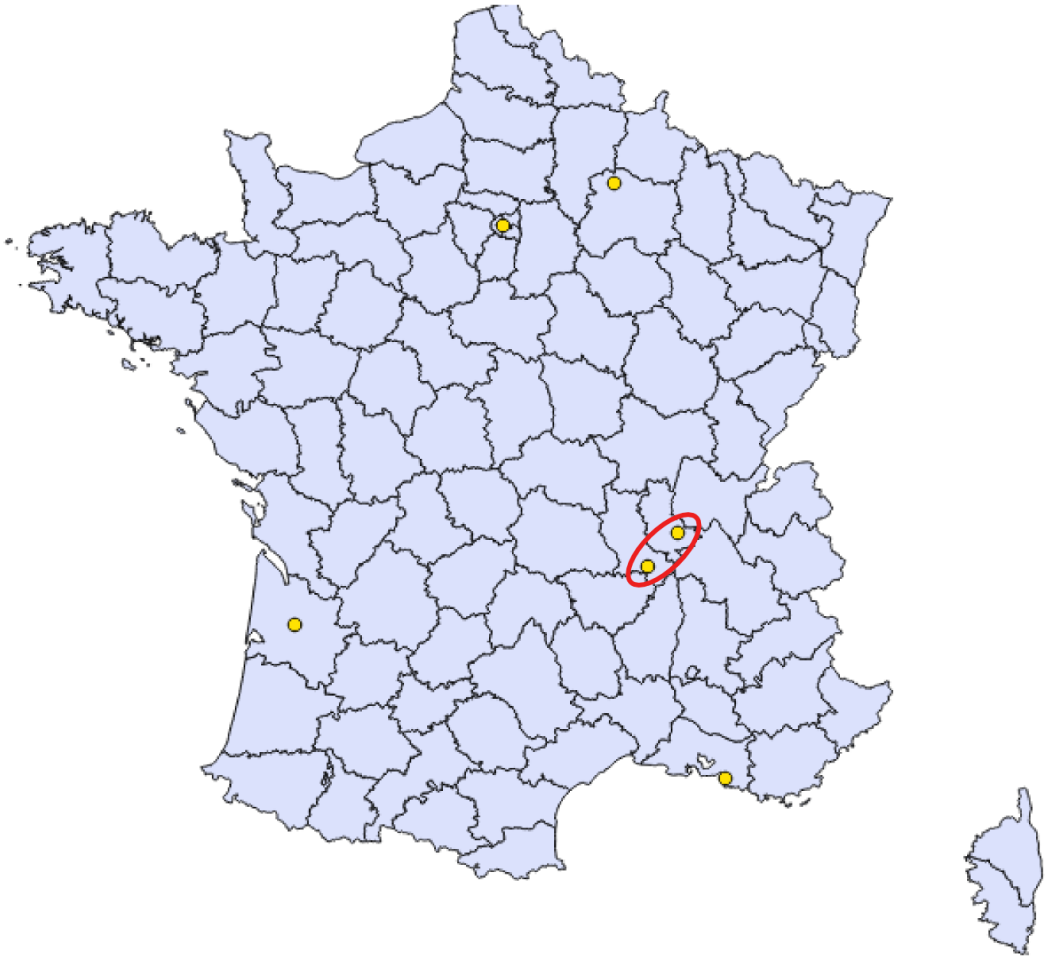
Principe

	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0

CAH

Principe

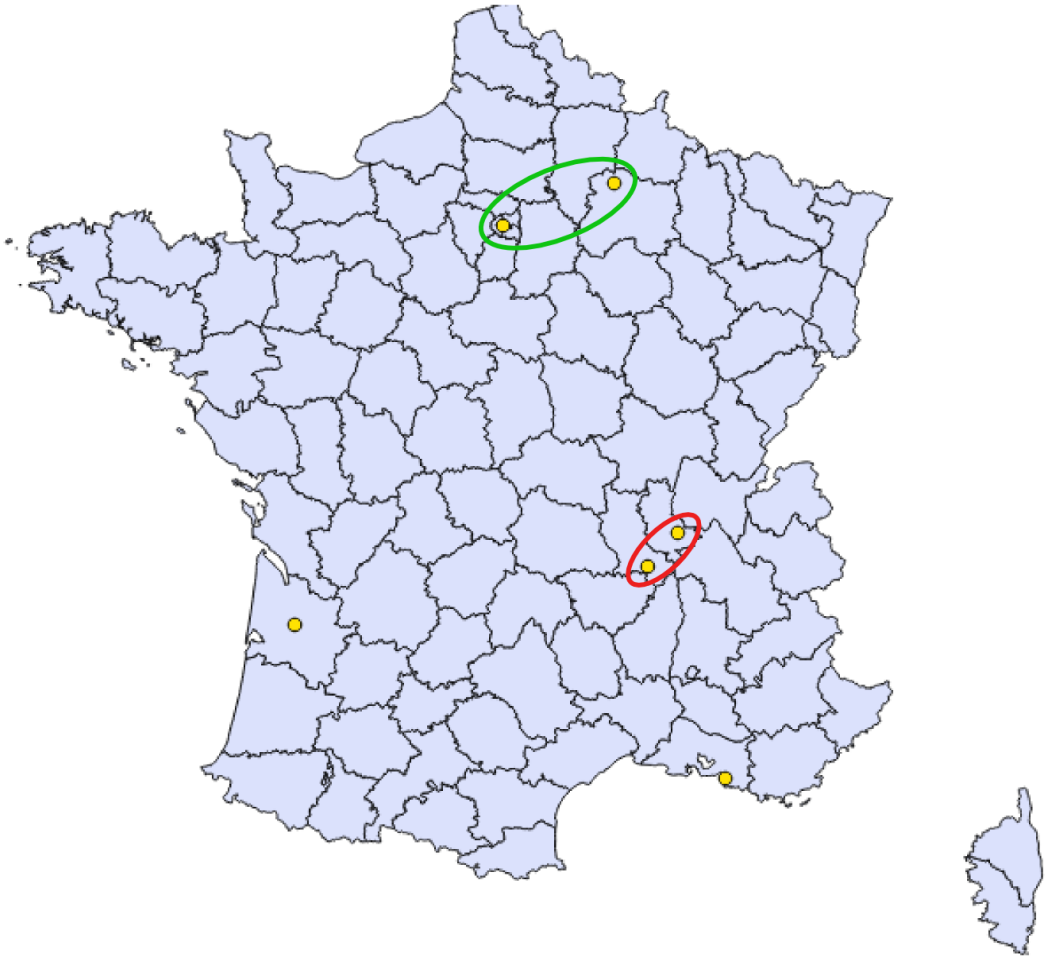
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



CAH

Principe

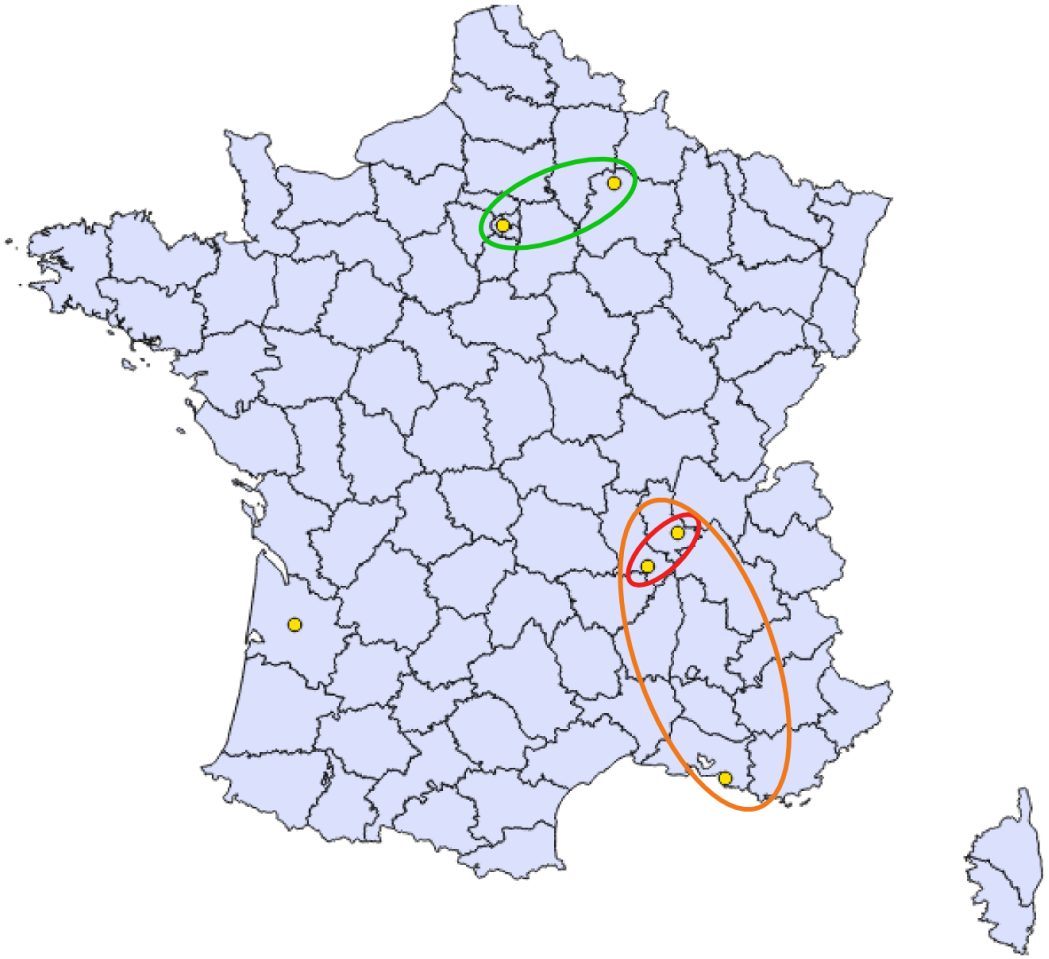
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



CAH

Principe

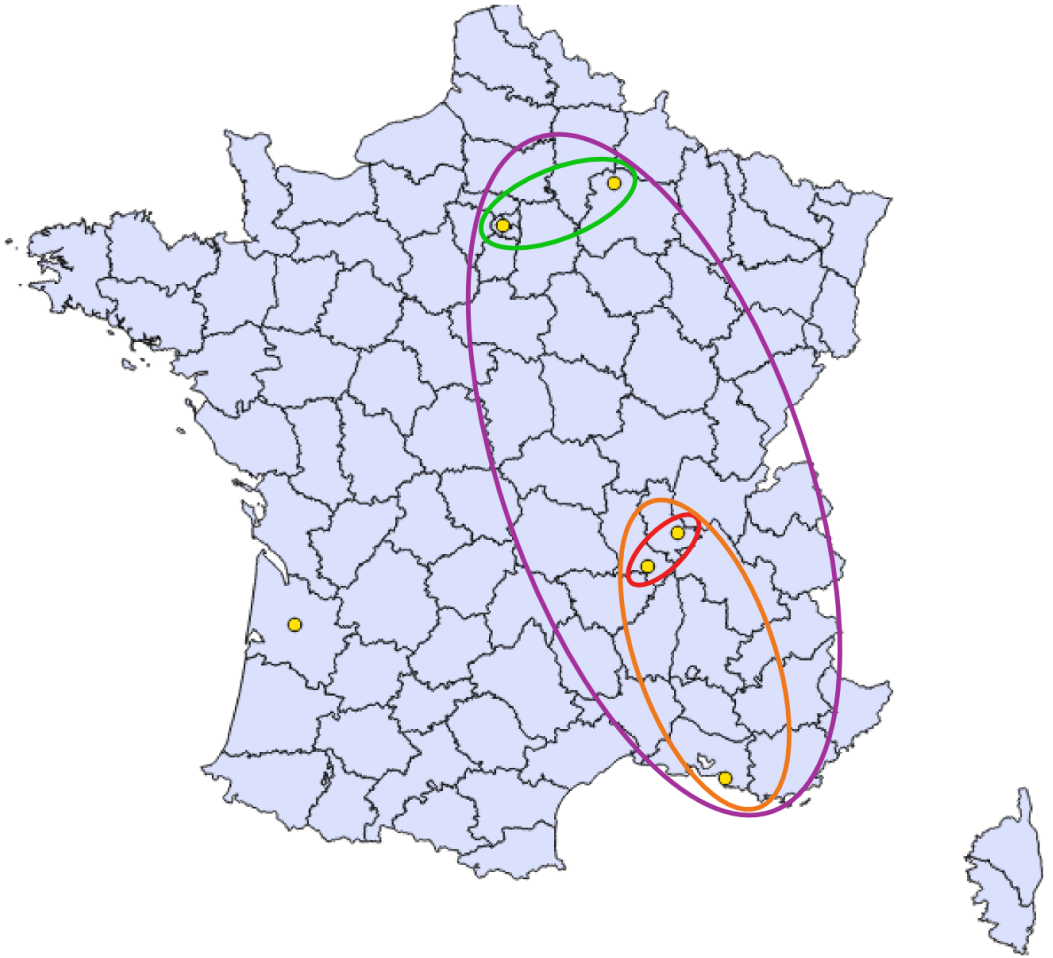
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



CAH

Principe

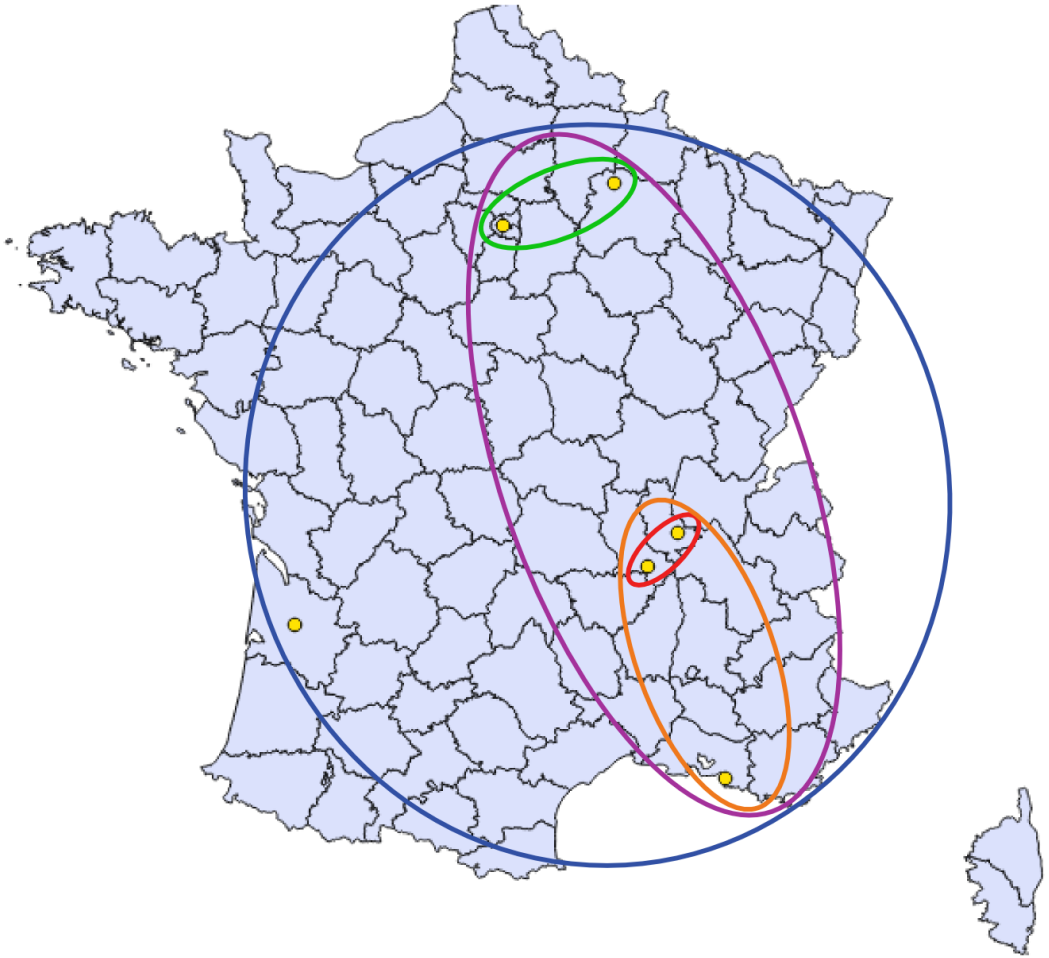
	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



CAH

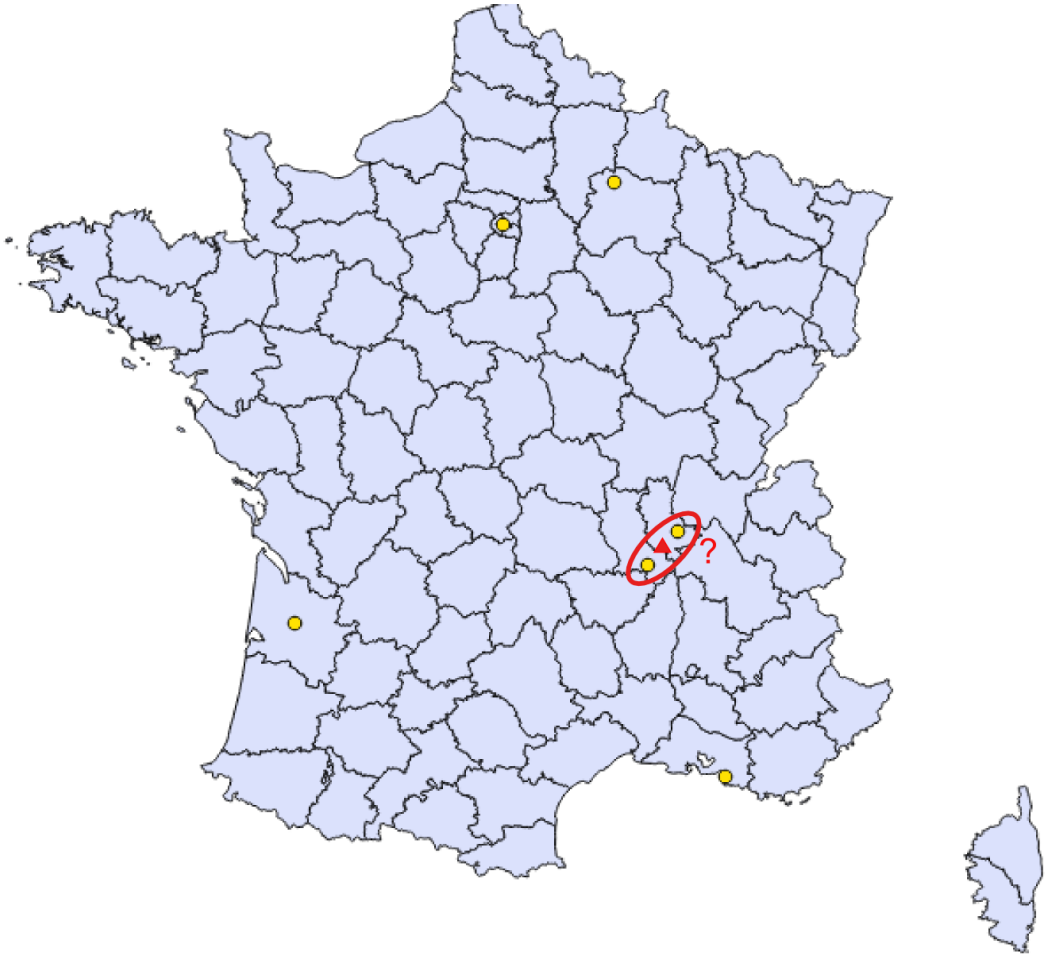
Principe

	Paris	Marseille	Saint-Etienne	Bordeaux	Reims	Lyon
Paris	0	660	410	498	131	392
Marseille	660	0	250	506	670	276
Saint-Etienne	410	250	0	396	425	50
Bordeaux	498	506	396	0	602	436
Reims	131	670	425	602	0	393
Lyon	392	276	50	436	393	0



CAH

Principe



CAH

Paramètres d'application

Parmi les paramètres d'une CAH, en plus de la mesure de dissimilarité, il y a donc le critère d'agrégation :

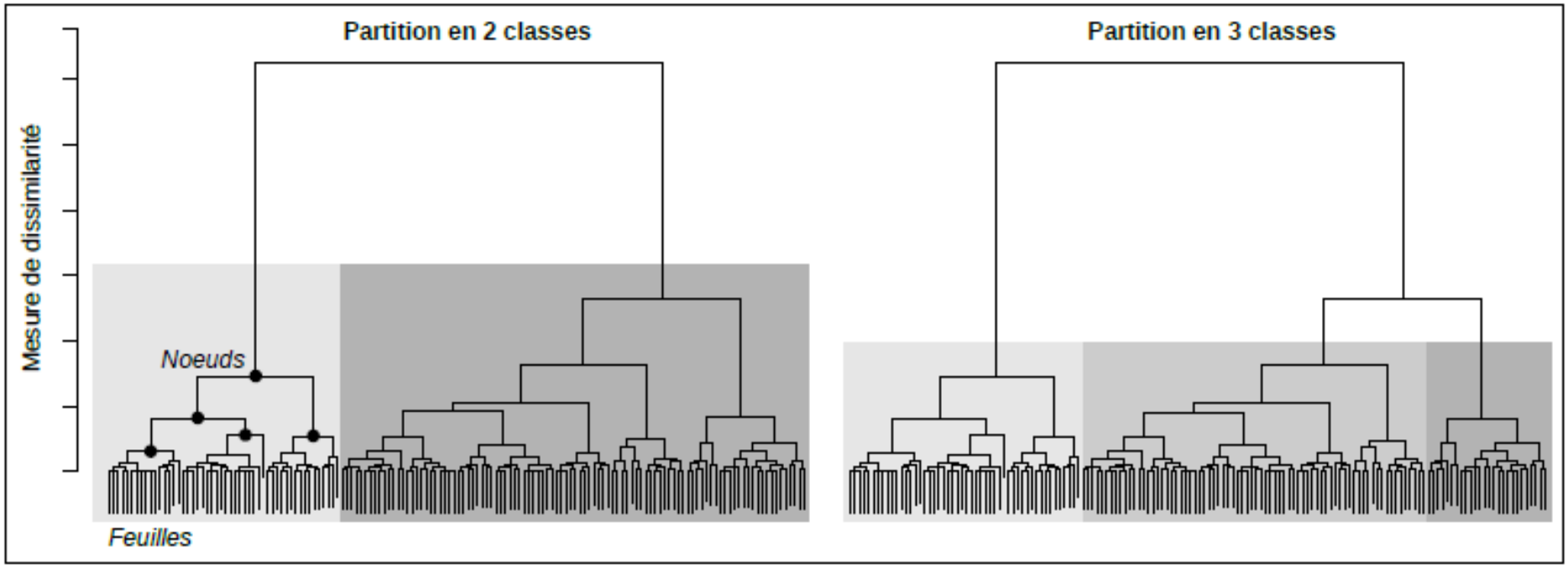
- Le saut minimum retient le minimum des distances entre individus de C1 et C2. C'est ce critère qu'on a appliqué précédemment.
- Le saut maximum est la dissimilarité entre les individus de C1 et C2 les plus éloignés.
- Le lien moyen consiste à calculer la moyenne des distances entre les individus de C1 et C2.
- La distance de Ward vise à maximiser l'inertie inter-classe.

Il faut aussi choisir le nombre de classes en tenant notamment compte de la qualité de la partition qui se mesure à l'aide d'une valeur d'inertie (variance).

Un gros travail consiste à interpréter les caractéristiques des classes créées.

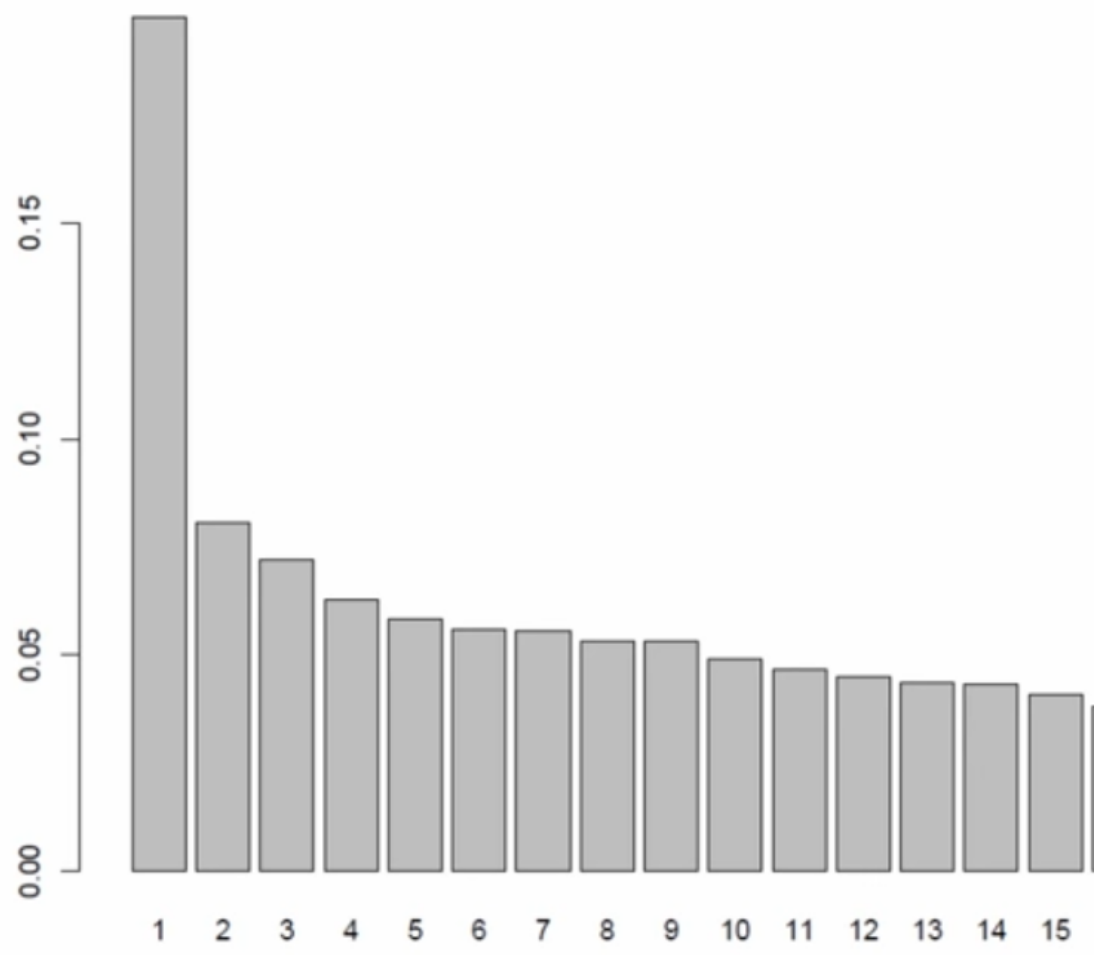
CAH

Dendogramme



Variance et Inertie

La qualité des axes principaux ou des classifications



1 Introduction

2 Corrélation et régression de variables quantitatives

3 Variables qualitatives

4 Méthodes exploratoires, synthétiques et classification

5 Interprétation des résultats obtenus sous R

6 Introduction à R

1 Introduction

2 Corrélation et régression de variables quantitatives

3 Variables qualitatives

4 Méthodes exploratoires, synthétiques et classification

5 Interprétation des résultats obtenus sous R

6 Introduction à R

Affectation et calcul

R fonctionne un peu comme une calculatrice. Si vous tapez $2 + 3$, le logiciel vous retournera la valeur 5. Néanmoins, on utilisera R davantage comme un langage de programmation en suivant les principes de l'affectation informatique.

Exemple d'affectation avec R

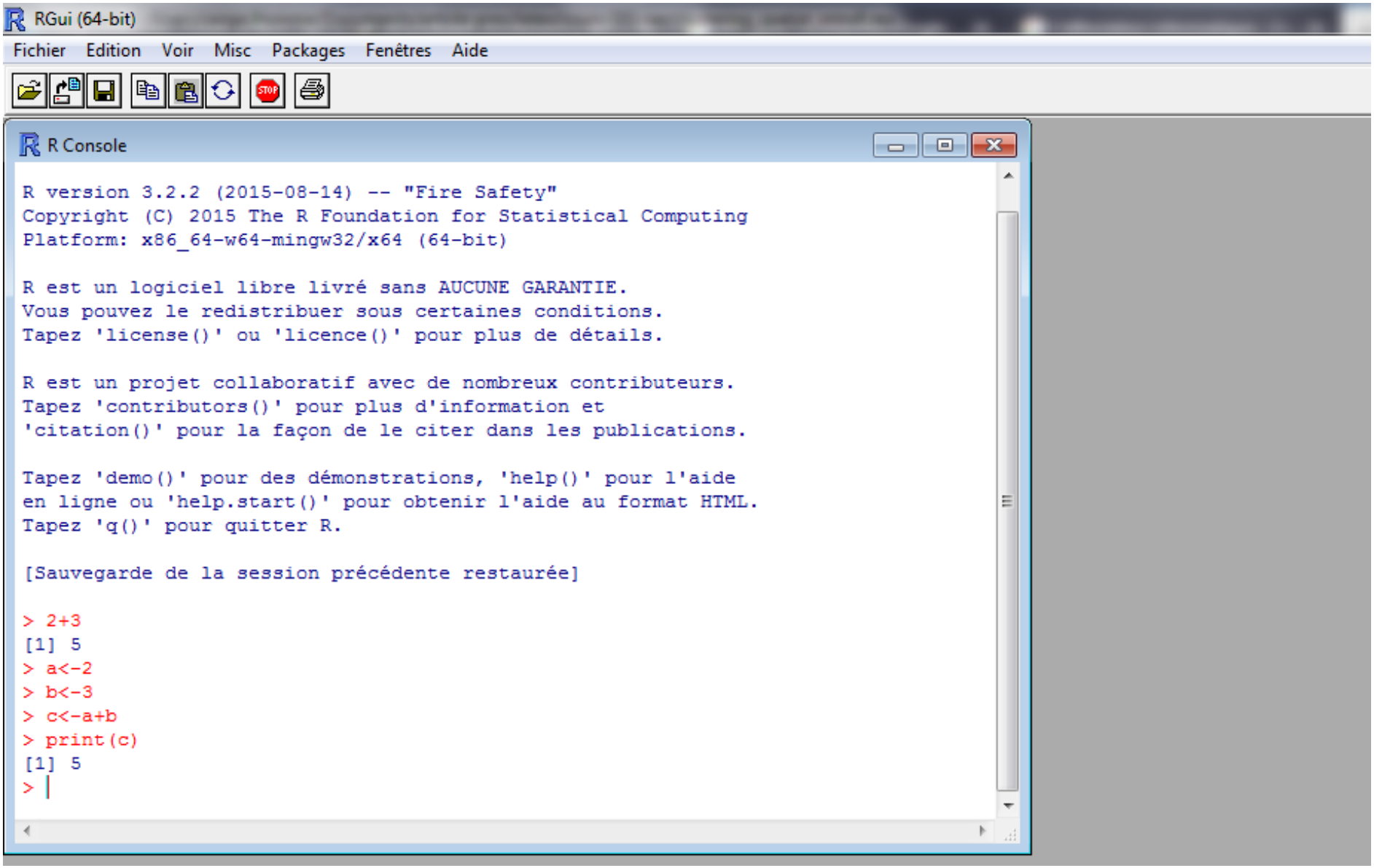
```
a <- 2
b <- 3
c <- a + b
```

L'affichage des résultats se fera alors en utilisant une fonction : « print() ».

Affichage d'une variable avec R

```
print(c)
```

Affectation et calcul



Les types de données

Il existe de nombreux types de variables dans R.

Les variables de type texte

```
a <- "Texte"
```

Ces variables peuvent être ordonnées dans une liste (un vecteur) ou dans plusieurs listes pour former une matrice (un tableau de valeurs).

Les vecteurs et les matrices

```
b <- c(18, 182, 1.5, 15, 200, 5)
c <- matrix(c (18, 182, 1.5, 15, 200, 5), nrow = 2)
d <- matrix(c (18, 182, 1.5, 15, 200, 5), ncol = 2)
```

Les types de données

Pour accéder à une valeur ou à un ensemble de valeurs, il faut utiliser les index des vecteurs ou des matrices.

Accès aux valeurs des vecteurs et des matrices

```
e <- b[2] + b[3]
f <- c[1,2] + c[2,3]
col <- c[,1]
ligne <- c[1,]
```

Accès avancé aux valeurs des vecteurs et des matrices

```
e <- b[c(2,4)]
f <- c[(c<15)]
g <- b[2 :5]
```

Les types de données

Les data frames permettent de manipuler des tableaux bien structurés. Ce type de données est particulièrement bien adapté aux importations de fichiers textes.

Les Data Frames

```
articles <- c( "un", "le", "la", "les")
sujets <- c( "mot", "terme", "chose", "images")
dfmots <- data.frame(articles, sujets)
dfmots2 <- data.frame(col1 = articles, col2 = sujets)
```

Appel des valeurs des Data Frames

```
print(dfmots$sujets)
print(dfmots[,1])
```

L'import de données et premières fonctions

Importation de fichiers textes

```
MyTexte <- read.table(file="c :/TheData.csv", header=TRUE, sep=",")
MyData <- read.csv(file="c :/TheData.csv", header=TRUE, sep=",")
adresse <- file.choose()
MyData <- read.csv(file=adresse, header=TRUE, sep=",")
```

Fonctions de base

```
res <- summary(b)
plot(d[,1],d[,2])
hist(b)
reg <- lm(d[,1] ~d[,2])
res3 <- summary(reg)
t.test(d[,1], d[,2])
```

Les bibliothèques

Ce qui constitue la puissance de R, ce sont ses nombreuses bibliothèques qu'il faut télécharger.

Les librairies cartographiques

```
library(rgdal)
nuts3 <- readOGR(dsn = adresse, layer = "nuts3", verbose = TRUE)
library(sp)
class(nuts3)
nuts3@proj4string
head(nuts3@data)
plot(nuts3[1, ], col = "#5C99AD", border = " #2A5F70", lwd = 4)
library(rgeos)
europeBuffer <- gBuffer(spgeom = europe, width = 50000)
```

Les boucles

Enfin, comme tout langage de programmation, R permet de répéter les mêmes instructions plusieurs fois en changeant seulement quelques paramètres. Ce sont les boucles. Ces boucles peuvent alors permettre d'effectuer des tests. Ce sont par exemple les Si.

Les boucles

```
for (i in 1 :10) {
        print(i)
        }
for (i in 1 :10) {
        if (i > 5 & i < 8) {
                print(i)
                }
        }
```