Raisonner toutes choses égales par ailleurs?

Partie 1. L’ambition d’un raisonnement toute chose égale par ailleurs
Partie 2. Ceux qui raisonnent toutes choses inégales par ailleurs

Partie 1. L’ambition d’un raisonnement toute chose égale par ailleurs

Marshall (1890)

« On dit parfois que les lois de l'économie sont « hypothétiques ». Bien sûr, comme toute autre science, l'économie entreprend l'étude des effets que produisent certaines causes, non pas absolument mais **sous la condition que toute chose soit égale par ailleurs** et que les causes puissent produire leurs effets **sans interférence** (…) L'action des causes en question est supposée **isolée**; certains effets leur sont attribués, mais seulement sous l'hypothèse qu'aucune autre cause n'est autorisée à intervenir, en dehors de celles qui ont été directement permises. »

1. neutraliser les effets de structure
2. Construire des régressions linéaires
3. Construire des régressions logistiques
4. Utiliser une variable instrumentales
5. Construire un expérience contrôlée
6. utiliser une expérience naturelle

1. Neutraliser les effets de structure

Repérer un effet de structure

Catégories	1982	2009
Diplômés	8,6	8,9
Non diplômés	13,6	13,9
Population totale	12,35	11,9

(8.6 + 13.6)/2 = 11.1 = moyenne simple des taux de chômage par catégorie

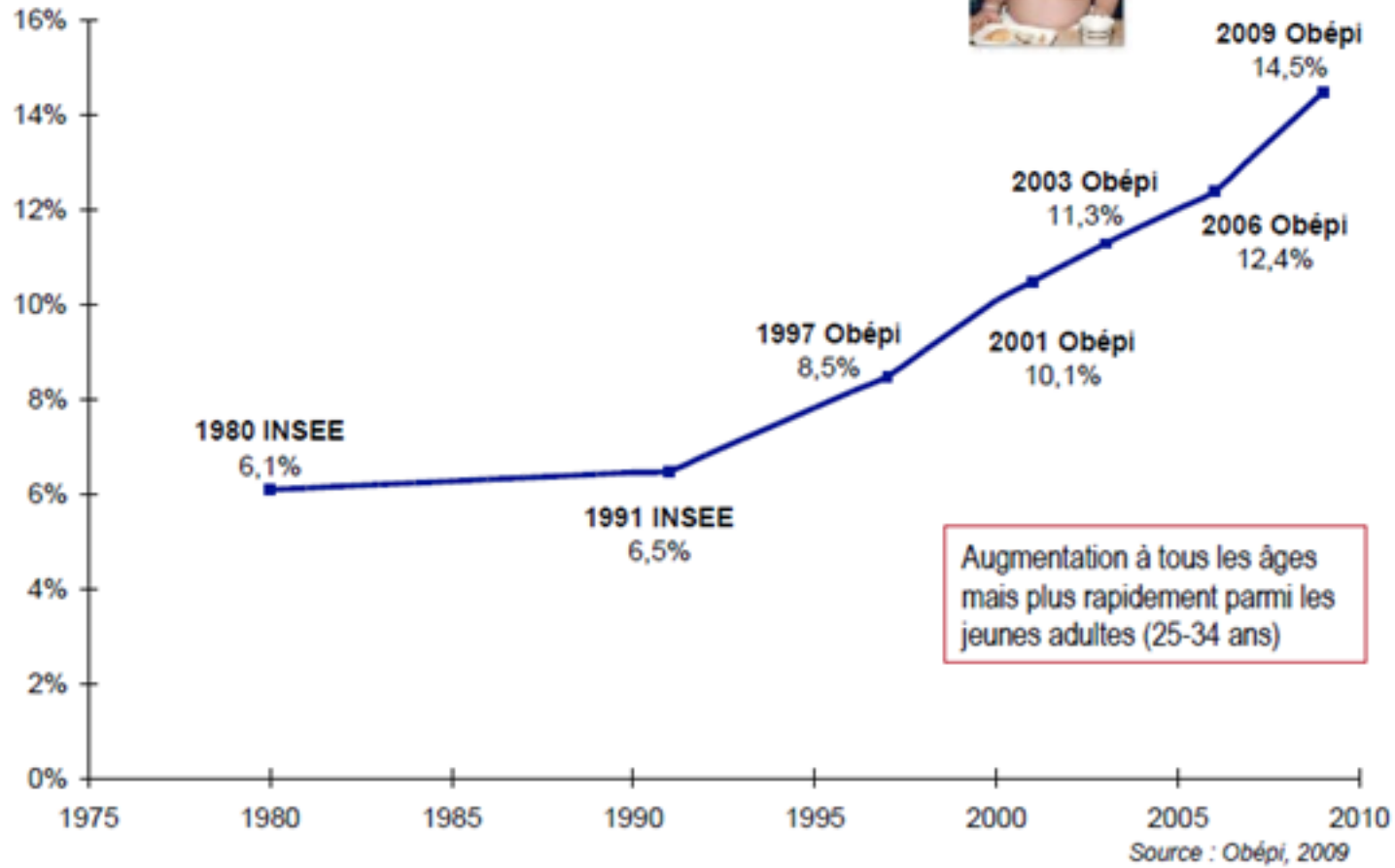
12.35, c'est la moyenne pondérée par le poids des deux catégories dans la population en 1982.

Le taux de chômage augmente dans les deux catégories et pourtant il baisse dans l'ensemble de la population sur la même période.

Conclusion : la part des diplômés dans la population active a augmenté, leur poids dans la moyenne pondérée aussi, et comme leur taux de chômage est plus faible, ces diplomés font baisser la moyenne

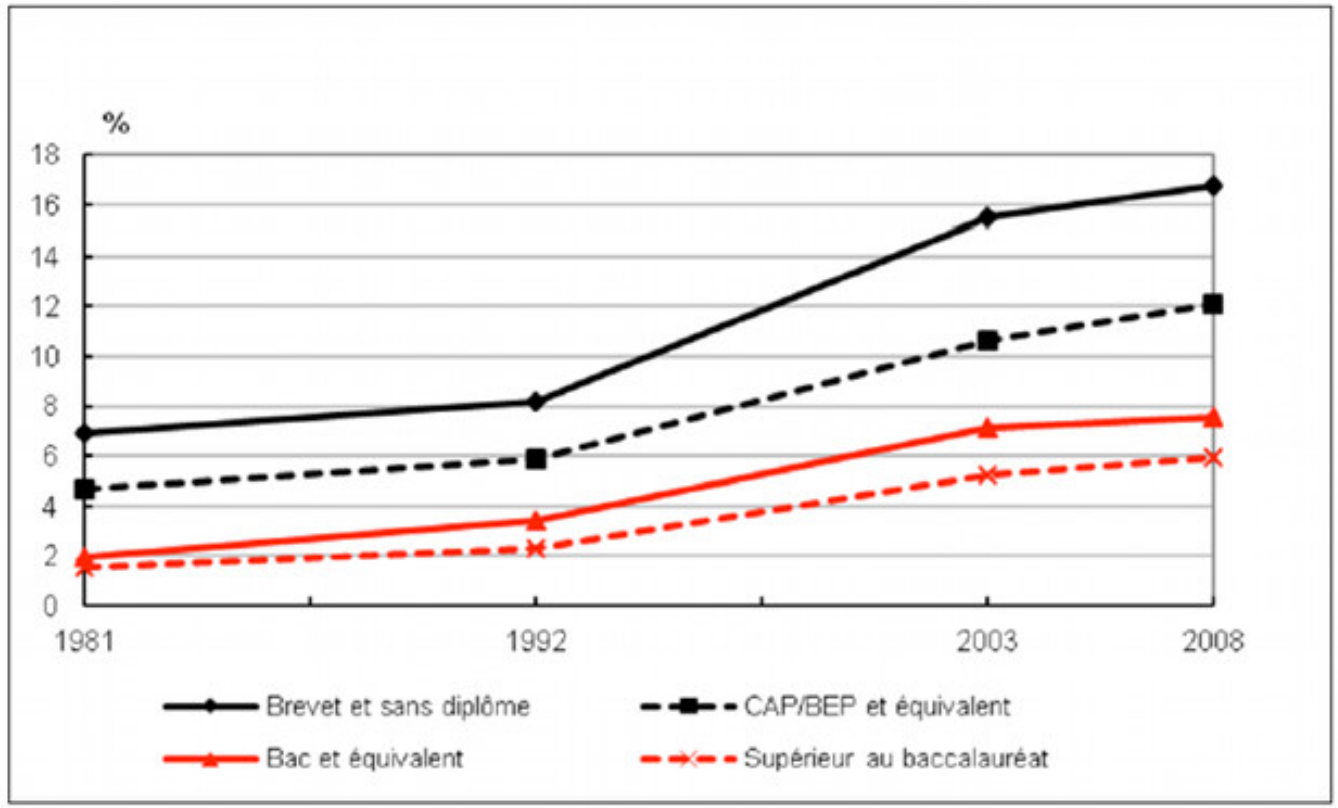
Augmentation de l'obésité en France

Prévalence de l'obésité chez les adultes de 15 ans et plus



Source : Obépi, 2009

Figure 1 : Evolution de la prévalence de l'obésité selon le niveau de diplôme



Lecture : en 1981, 6,9 % des individus sans diplôme ou ayant au plus un brevet étaient obèses.

Source : INSEE, enquêtes Santé.

Repérer un effet de structure

Par exemple, le salaire de chaque profession peut stagner (ou augmenter faiblement) alors que le salaire moyen augmente fortement.

On observe ce phénomène quand les professions très qualifiées et donc mieux payées voient leur poids augmenter dans la population active.

On peut neutraliser cet effet de structure, en calculant la variation qu'on aurait observée à structure de la population constante, on l'appelle variation nette.

Variation totale = variation due à l'effet de structure + variation nette

un biais dans la lecture et l’interprétation

- des évolutions des pratiques ou des représentation d’une population entre deux dates
- de la comparaison entre de deux populations différenciées à une même date

On peut neutraliser les effets de structure

On cherche à distinguer

- Dans une évolution : ce qui s'explique par l'évolution de la structure de la population (l'évolution du poids des différentes catégories dans cette population) et ce qui s'explique par une évolution des pratiques ou des représentations dans les différentes catégories de la population.
- Dans une comparaison : ce qui s'explique par le fait que dans les deux populations la répartition par catégories est différente et ce qui s'explique autrement.
- Dans les deux cas, on décompose en un effet structurel et un effet net les différences observées (entre deux dates ou entre deux populations)

Un exemple d'effet de la structure par âge

- On voudrait savoir si le fait d'être sorti jeune du système scolaire augmente la probabilité de se trouver en prison.
- Pour ça, on choisit de comparer les âges de sortie du système scolaire dans les deux populations.
- On observe que les hommes qui ont arrêté jeune leurs études sont surreprésentés en prison (le pourcentage de ceux qui sont sortis tôt du système scolaire est supérieur à celui observé dans le reste de la population)
- Mais on se heurte à une difficulté d'interprétation de ce constat

- On observe que la population des hommes détenus est plus jeune que la population des hommes libres : leur structure par âge est différente
- On sait par ailleurs que la durée des études a progressé et que donc la probabilité d'avoir arrêté tôt ses études est plus forte pour les hommes qui appartiennent à des catégories d'âge plus vieilles
- On ne peut donc pas comparer les deux populations sans tenir compte de leurs différences de structure par âge, puisque cette structure par âge influence la probabilité d'avoir arrêté tôt ses études.

Document 1. Structure par âge de la population carcérale et de la population des hommes libres (%en ligne)

	18-24	25-29	30-34	35-39	40-44	45-49	50-59	+de 60
Part relative chez les détenus	20.4	18.4	14.4	12.6	10.5	9	10.4	4.4
Part relative chez les hommes « libres »	9.6	9.5	9.9	10.3	10	10.1	15.3	24.1

Document 2
Age de sortie du système scolaire chez les hommes en liberté (en %)

	18-24	25-29	30-34	35-39	40-44	45-49	50-59	60 et +	Ensemble	**Ensemble à âge comparable**
15 ans et avant	3.5	3.1	4.7	6	12	27.1	37.3	58.4	26.4	12.8
Entre 16 et 17 ans	19.4	17.4	25.1	32.1	33.9	22.9	18.9	13.8	21.5	23.1
Entre 18 et 19 ans	29.9	24.3	28.5	29.4	24.5	20.2	15.9	9.5	20.2	24.8
Entre 20 et 24 ans	33.2	43.8	29.6	21.1	19.2	18.6	16.3	8.3	20.6	27.5
25 ans et plus		7	7.7	7.1	6	6.6	6.4	3.6	5.5	5.4

Source : INSEE, EHF 1999 auprès des hommes en ménage ordinaire

Comparaison des âges de sortie du système scolaire

	15 ans ou avant	Entre 16 et 17 ans	Entre 18 et 19 ans	Entre 20 et 24 ans	25 ans et plus
Hommes détenus	27.7	44.3	18.2	8.3	1.5
Hommes libres	26.4	21.5	20.2	20.6	5.5
Hommes libres « à âge comparable »	12.8	23.1	24.8	27.5	5.4

La structure par âge est différente dans les deux populations

Pour neutraliser cet effet de structure, on doit pouvoir raisonner « à âge égal », et reconstituer par le calcul ce que serait l'âge de sortie du système scolaire des hommes libres s'ils avaient la même structure par âge que la population des hommes détenus.

C'est le résultat de ce calcul qu'on trouve dans la dernière colonne du document 2 et repris dans la dernière ligne du tableau de la slide précédente

Proportion d’hommes libres sortis du système scolaire à 15 ou avant qu’on observerait si les structures par âge des deux populations étaient les mêmes

=
3.5(0.204)+3.1(0.184)+4.7(0.144)+6(0.126)+12(0.105)+27.1(0.09)+37.3(0.104)

+58.4(0.044)

=12.865

12.8%, c’est bien la valeur qu’on lit sur la première ligne dans la colonne « à âge comparable » du document 2

Si on raisonne à âge comparable

- La probabilité d'avoir quitté tôt le système scolaire diminue pour les hommes libres
- La probabilité d'avoir continué ses études augmente pour les hommes libres
- C'est ce qui permet de conclure que les hommes qui ont eu une scolarité plus courte sont bien surreprésentés dans la population carcérale = toutes choses égales par ailleurs, les hommes qui sortis jeunes du système scolaire ont une probabilité plus forte de se trouver en prison.

La réussite au bac est-elle supérieure chez les filles?

Dans une ville imaginaire

	effectifs	succès	échec	% de réussite
garçons	60	24	36	40%
filles	60	36	24	60%

Les résultats dans les deux lycées de la ville

	effectifs	succès	échec	% de réussite
garçons	50	15	35	30%
filles	10	1	9	10%

	effectifs	succès	échec	% de réussite
garçons	10	9	1	90%
filles	50	35	15	70%

Interprétation des résultats

- Les deux lycées n'ont pas la même structure par genre des élèves.
- Les filles sont surreprésentées dans le lycée où les taux de réussite sont les meilleurs.

Pour interpréter les résultats bruts de la première comparaison (filles/garçons), on a besoin d'introduire des variables supplémentaires = des variables test ou des variables de contrôle = ici une meilleure connaissance des établissements de la ville permet d'affiner l'analyse et d'introduire la variable de l'établissement comme variable test, pour vérifier si les filles réussissent mieux toutes choses égales par ailleurs le bac.

2. Construction et lecture des régressions

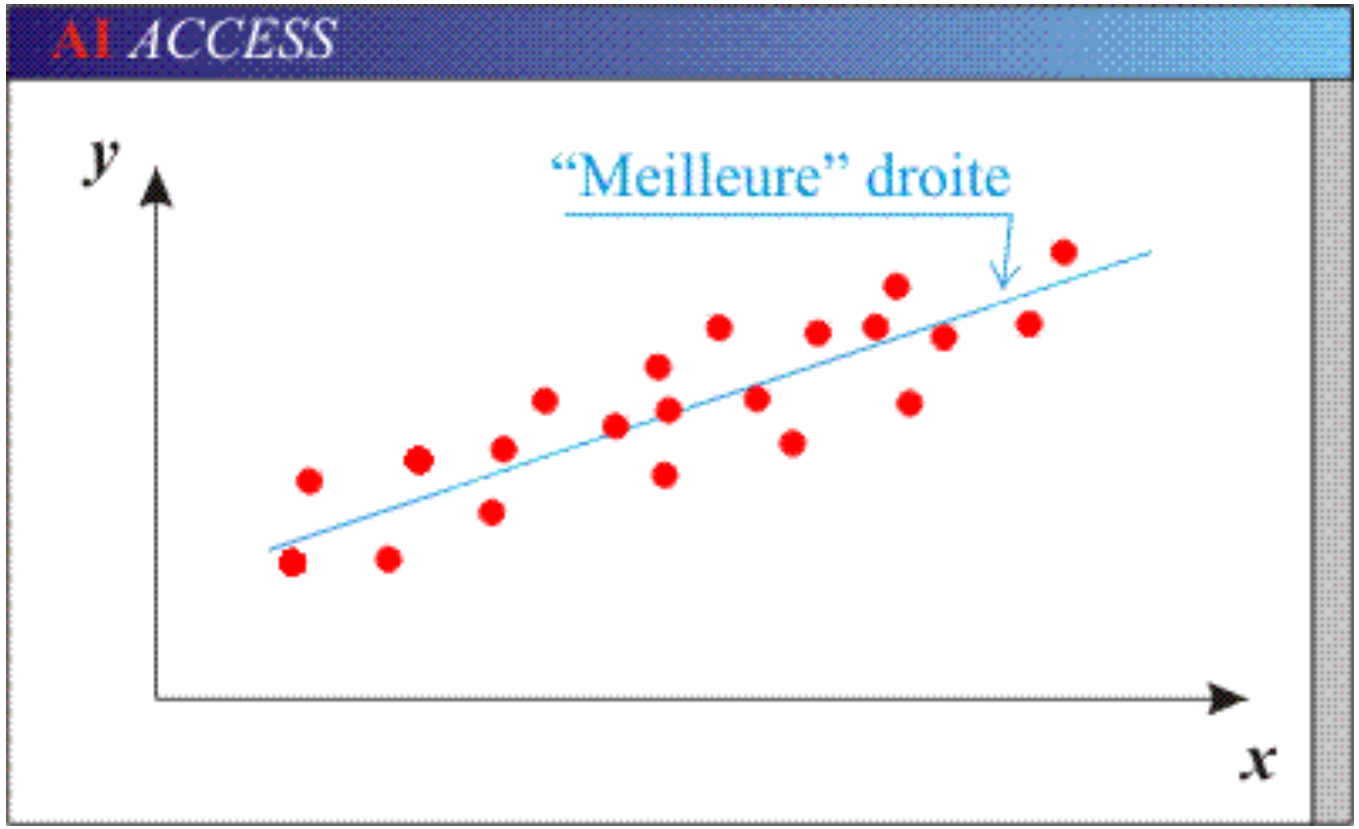
Lire une régression

La régression linéaire simple :

Objectif : on cherche à expliquer les variations d'une variable quantitative par les variations d'une autre variable quantitative.

- La variable à expliquer est notée Y : **variable dépendante** – variable de réponse
- La variable explicative est notée X : **variable indépendante**

La droite de régression



Elle permet de visualiser une **corrélation** entre les deux variables et de la caractériser. Le **coefficient de régression** correspond à la pente de la droite. On peut déterminer si il est **positif ou négatif** et aussi sa **valeur absolue**.

L’interprétation du coefficient = un exemple

SAL = a + b.EDU + ϵ

Ici SAL c’est le salaire mensuel en euros = la variable dépendante
EDU c’est le nombre d’années d’études = la variable indépendante
La relation entre le salaire et le nombre d’années d’études c’est la relation d’intérêt
b = on l’appelle le paramètre d’intérêt
$\in$ = le terme d’erreur = il intègre les aléas liés au tirage d’un échantillon particulier et les variables omises.

Si b estimé (noté b^) à partir des observations dans l’échantillon = 400. souvent on note entre parenthèse l’écart-type par exemple (50)

On obtient la droite de régression par la méthode des moindre carrés, pour tous les points de l’échantillon, on cherche à minimiser l’écart entre son ordonnée et l’ordonnée estimée par la droite.

Si b = 400, on peut dire qu’une année de scolarité en plus augmente statistiquement de 400 euros le salaire mensuel.
Mais il faut prendre quelques précautions

tester la qualité de la régression

On peut d'abord calculer un **coefficient de détermination, noté R^2** , il mesure l'adéquation de la droite au nuage de points. Il mesure en fait la part de la variance totale de Y qui est expliquée par la régression qu'on a construit.

A retenir : Il varie entre 0 et 1 ou entre 0% et 100% si il est exprimé en pourcentage. **Logiquement plus on se rapproche de 1 ou de 100% et meilleure est la régression**

La significativité des coefficients

- Attention, R2 ne permet pas de savoir si le modèle est statistiquement pertinent pour expliquer les valeurs de Y. Pour le déterminer, il faut utiliser des tests d'hypothèses pour vérifier si la liaison mise en évidence avec la régression n'est pas un simple artefact, lié à une échantillon particulier.
- Principe : on calcule la probabilité que le coefficient de régression soit nul, si cette probabilité est trop forte on considère que le coefficient qu'on a trouvé n'est pas significatif.
- Lecture : le seuil de signification du coefficient est normalement indiqué, souvent avec un système d'*.

***, le coefficient est significatif au seuil de 1% (la probabilité pour le coefficient soit nul est inférieure à 1%)

**, le coefficient est significatif au seuil de 5%

*, le coefficient est significatif au seuil de 10%

La significativité des coefficients

- Intuitivement c’est la faible dispersion des points autour de la droite qui importe
- Si le coefficient trouvé est élevé en valeur absolue et que l’écart type est faible, il y a une faible probabilité qu’avec un autre échantillon on puisse trouver un coefficient nul.

La significativité des coefficients

- l’intervalle de confiance à 95% = la probabilité que le coefficient se trouve dans cet intervalle est de 95% = il y a 5% seulement de chance qu’en choisissant un autre échantillon on puisse trouver un coefficient qui ne soit pas dans cet intervalle
- On utilise parfois une règle un peu simplifiée pour déterminer cet intervalle de confiance = (b – 2 écart-type; b+ 2 écart-type)

Le t de Student

- On utilise parfois aussi un autre indicateur de la significativité mais qui est équivalent
- t = valeur absolue du coefficient/écart-type
- Puis on dit que si t est supérieur à 2 alors le coefficient est statistiquement significatif au seuil de 5%

La significativité statistique ne suffit pas à conclure

- On se heurte à la possibilité d'une variable omise ou cachée comme à chaque fois qu'on établit une corrélation statistique
- Il y a plusieurs façons de traiter cette difficulté :
- la régression multiple = on introduit des variables de contrôle (= variables test) pour essayer de raisonner vraiment toutes choses égales par ailleurs
- on construit une régression logistique
- on choisit une variable instrumentale
- on construit une expérience contrôlée
- on cherche une expérience naturelle

La régression linéaire multiple

Objectif : ici on cherche à mesurer les relations entre

- une variable à expliquer notée Y : variable dépendante – variable de réponse
- des variables explicatives notée X_i : variables indépendantes ou régresseurs. On distingue la variable d'intérêt et les variables de contrôle.

L’équation de la régression

C’est logiquement l’équation d’un espace à k dimensions

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Avec Xi de i=1 à i=K, qui sont les K variables explicatives
Y qui est la variable dépendante
Les βi s’appellent des coefficients de régression

Examiner la force de la corrélation mesurée par le modèle

- Pour mesurer la variance expliquée par le modèle, on utilise le coefficient de détermination multiple R2 égal au rapport entre la variance expliquée par l'ensemble des régresseurs et la variance totale de Y
- À retenir : si R2 est proche de 1, alors la part de la variance expliquée par le modèle est bonne. Si il est proche de 0, c’est le contraire
- Limite : R2 a tendance à mécaniquement augmenter à mesure que l'on ajoute des variables dans le modèle.
- si l'on veut comparer des modèles comportant un nombre différent de variables, on utilise le coefficient de détermination ajusté qui est corrigé des degrés de libertés.
- Attention, R2 ne permet pas de savoir si le modèle est statistiquement pertinent pour expliquer les valeurs de Y. Pour le déterminer, il faut utiliser des tests d'hypothèses pour vérifier si la liaison mise en évidence avec la régression n'est pas un simple artefact.

A retenir pour lire les résultats d'une régression linéaire

- Il faut examiner le ou les coefficients de régression : leur signe et leur valeur absolue
- Il faut être attentif au coefficient de détermination R^2 et savoir l'interpréter
- Il faut examiner quand ils sont indiqués les seuils auxquels les coefficients sont significatifs

3. Pourquoi des régressions logistiques?

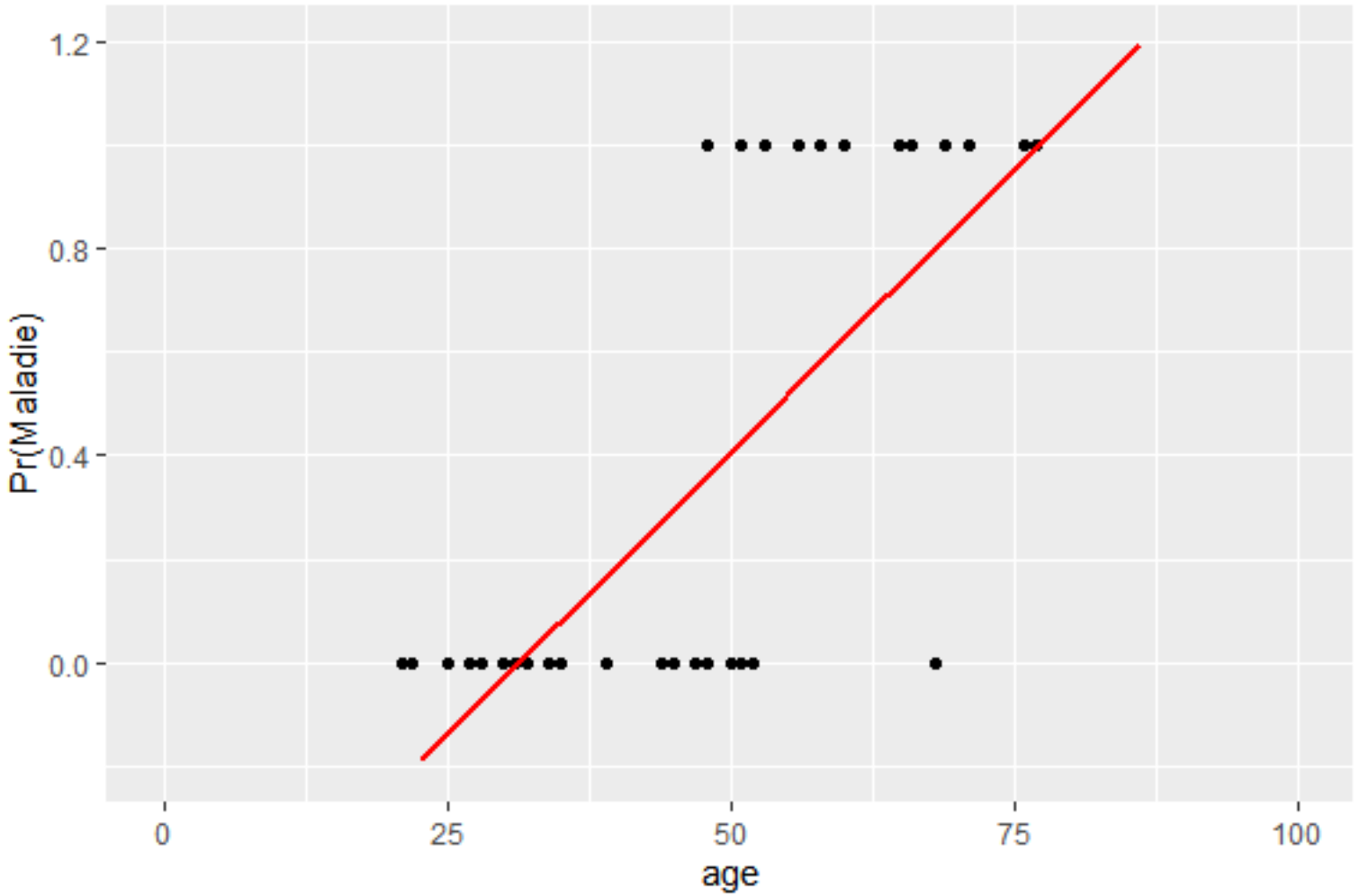
Le principe

- On cherche à modéliser, à partir d'un échantillon, les relations entre une variable dépendante et plusieurs variables indépendantes comme pour tous les modèles de régression
- La variable dépendante est dichotomique (exemple : voter Le Pen/ne pas voter Le Pen, voter/ne pas voter, être en surpoids/ne pas être en surpoids). On code Y = 1 ou Y = O.
- On voudrait isoler le poids de chaque variable indépendante en raisonnant « toutes choses égales par ailleurs ». Ces variables peuvent être qualitatives (souvent) ou quantitatives (mais souvent on les découpe en catégories)

Ce qu’on cherche à déterminer a la forme d’une probabilité conditionnelle

P(Y = 1/les caractéristiques d’un individu)

Un exemple médical : la probabilité que le patient soit porteur d'une maladie en fonction de son âge

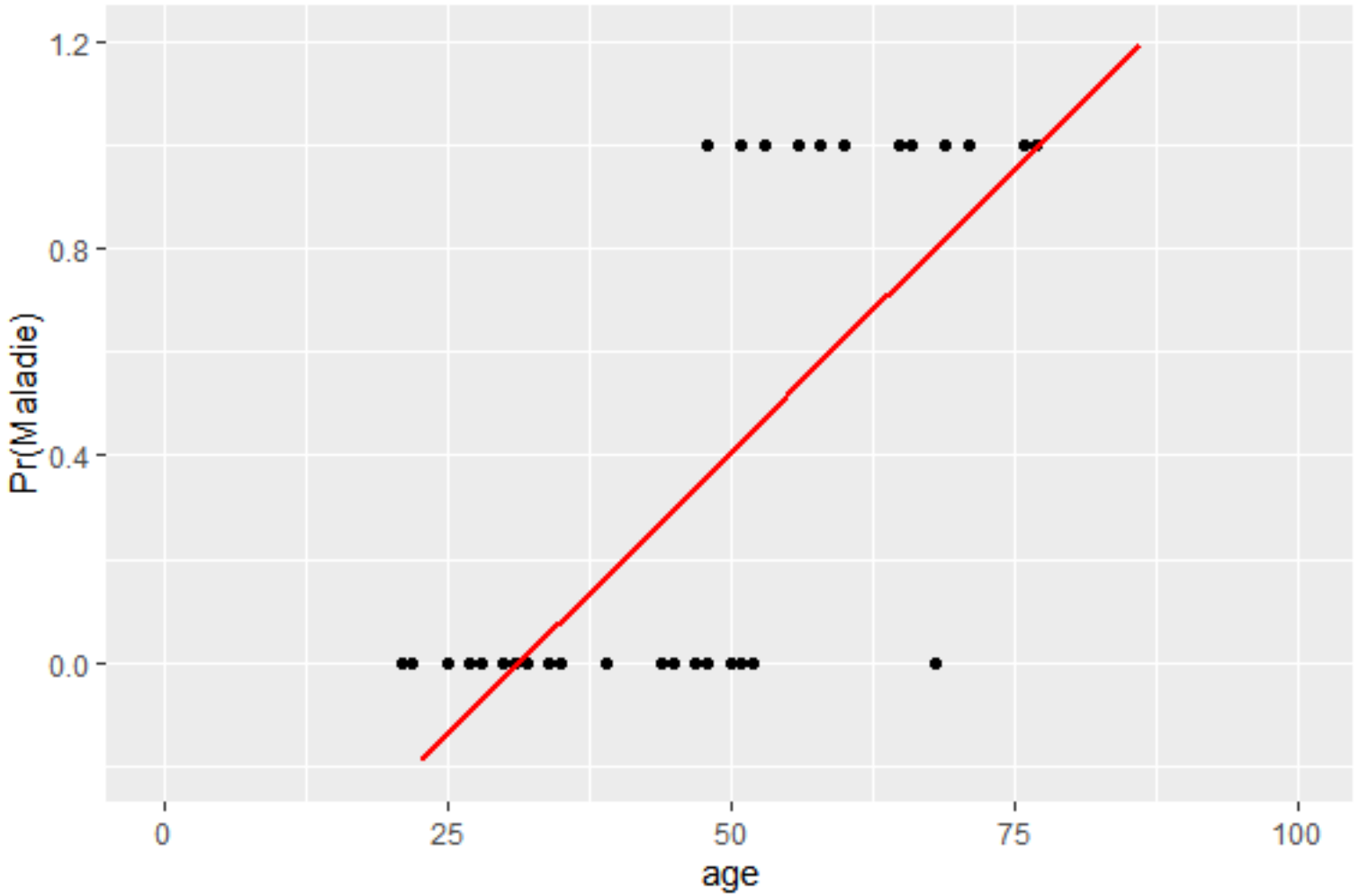


En abscisse
X = la variable explicative = la variable indépendante = l'âge. Elle peut prendre toutes les valeurs entre 0 et 100 (et plus)

En ordonnée

Y = être malade. Cette variable ne peut prendre que 2 valeurs : O et 1

On cherche P(Y=1/X)

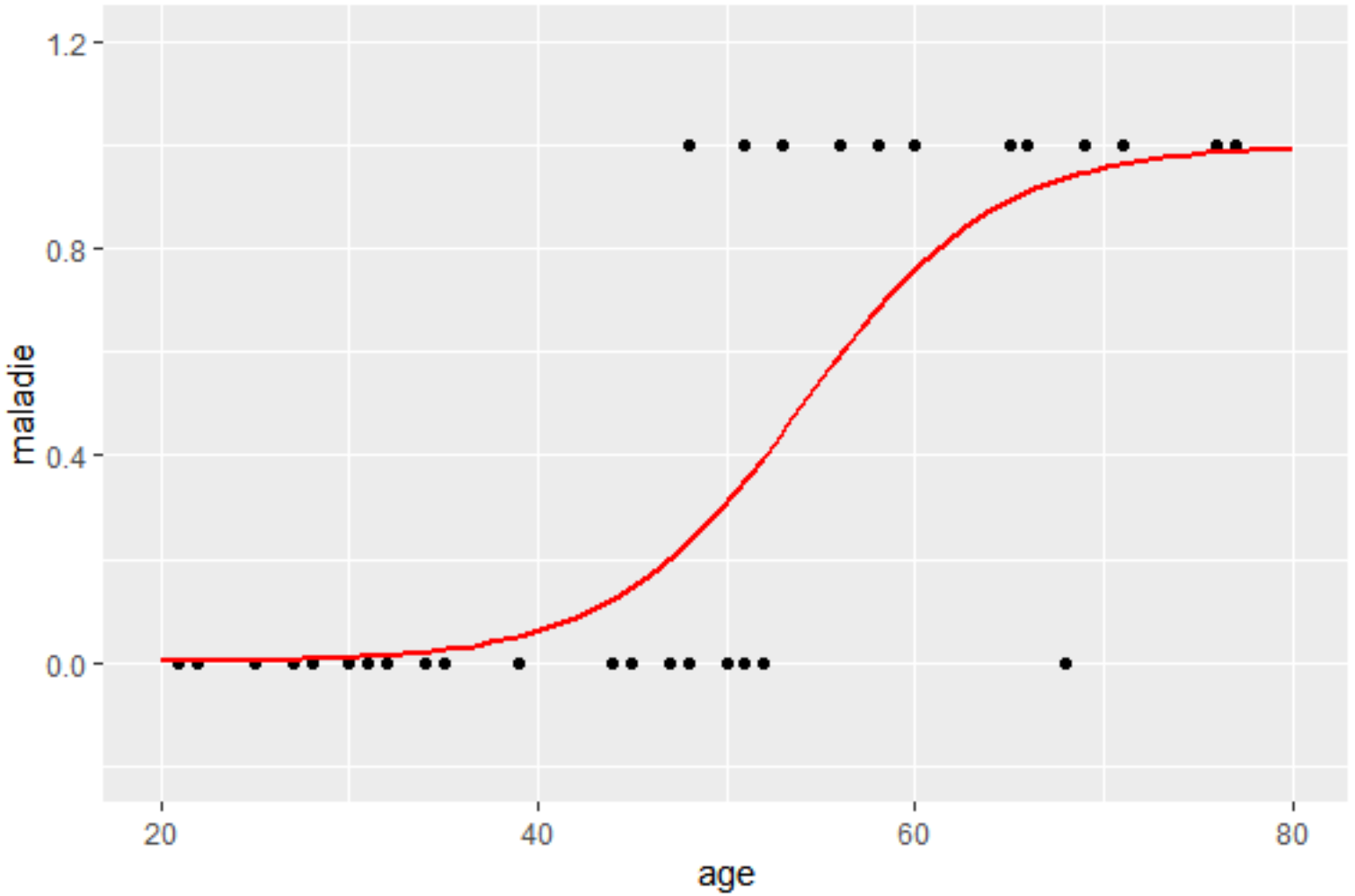


Si on construit une régression linéaire

On a une droite de régression d'équation

$P = aX + b$

Elle représente assez mal le nuage de points!!!



Ici la courbe est une sigmoïde, elle correspond à la courbe représentative de la fonction de répartition d'une loi logistique.

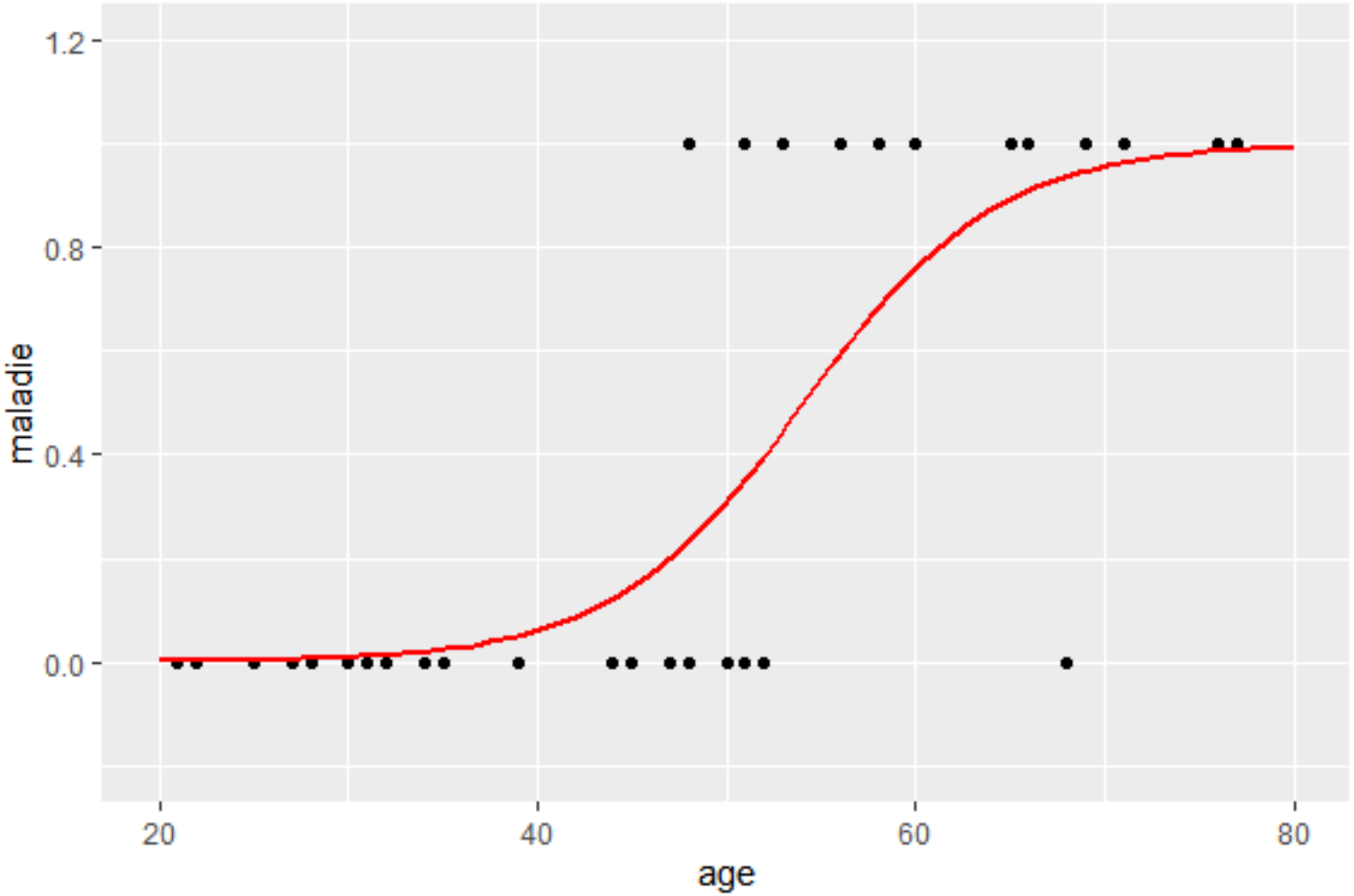
Cette fonction de répartition est de la forme

F(X) = P(X $< x$) =probabilité que la variable aléatoire X prenne une valeur inférieure à x =

$$= \frac{e^x}{1+ e^x}$$

Par définition, comme c'est une probabilité elle est comprise entre 0 et 1

Elle est définie pour toutes les valeurs possibles prises par x



P = P(Y/X) = C'est la probabilité d'être malade sachant X, l'âge qu'a le patient.

On peut chercher à estimer les paramètres (les régresseurs) $\beta 0\ et\ \beta 1, tels\ que$

P (Y/X) =

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

Ou, ce qui est équivalent :

$\beta_0 + \beta_1 X$ = ln (P/1-P) = logit P

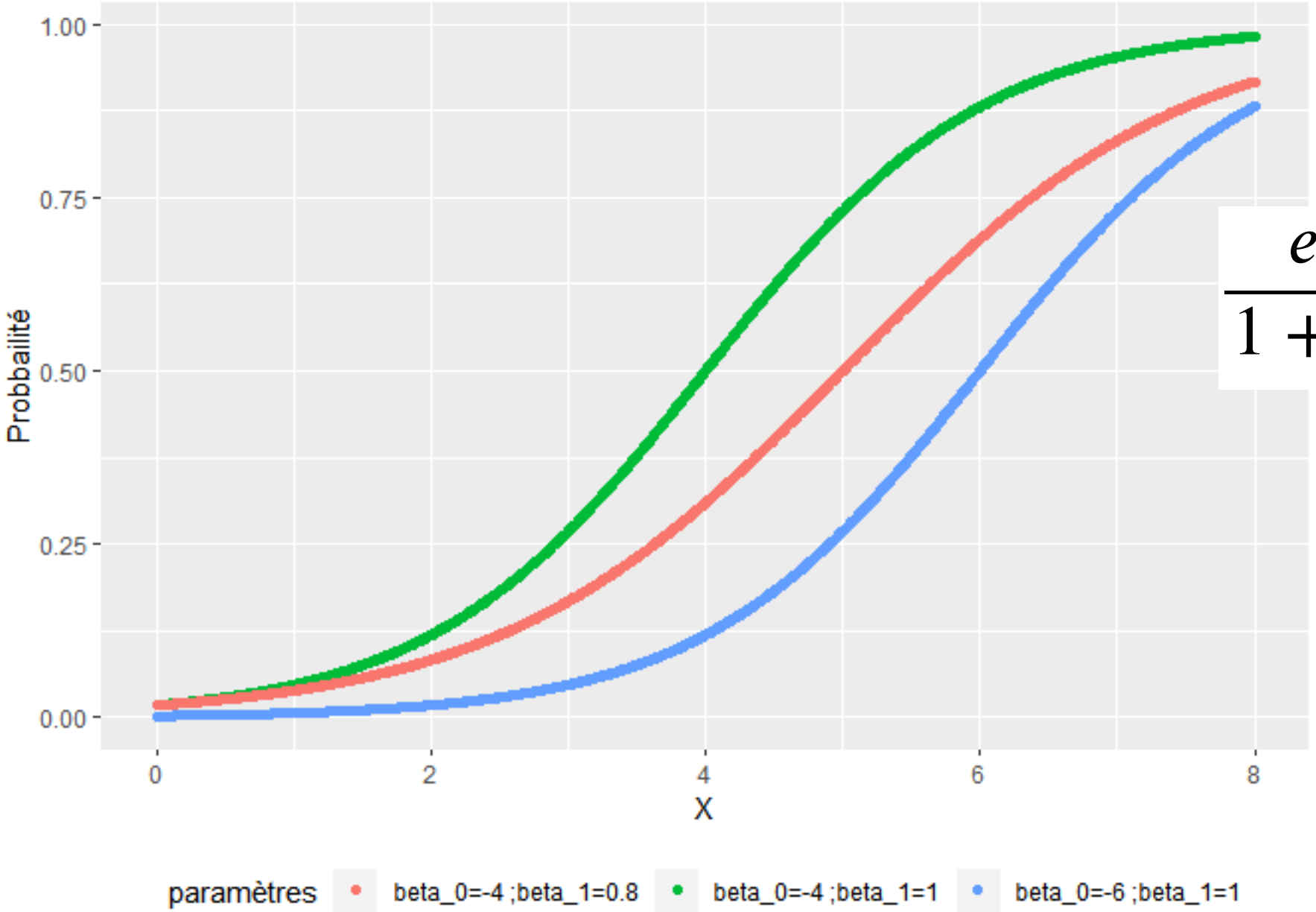
Rapport des chances relatives

Un **odds** en anglais c'est un rapport de chance, un rapport de deux probabilités : on compare la probabilité qu'un événement se produise avec la probabilité qu'il ne se produise pas. Il est de la forme :

$$\frac{p}{1-p}$$

Par exemple, probabilité d'être malade/probabilité de ne pas être malade

Un **odds-ratio**, c'est le rapport de deux odds, un rapport des chances relatives. On peut comparer les odds de deux sous-groupes d'un échantillon.



P(Y = 1/X l'âge du patient) =

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

La courbe varie en fonction de la valeur prise par les différents paramètres βi

On peut ensuite généraliser au cas de **plusieurs variables explicatives**

La variable Y est toujours dichotomique, on note Y = 1 ou Y = 0
Et P (Y = 1) = 1 – P(Y=0)

$$P(Y = 1/X_1, X_2, X_3, \ldots.Xn) = \frac{exp(\beta 0+\beta 1X1+\ldots+\beta nXn)}{1+exp(\beta 0+\beta 1X1+\ldots+\beta nXn)}$$

C'est la probabilité de développer une maladie (par exemple) sachant toute une série de caractéristiques qu'on a (âge, sexe, PCS, lieu de vie,….)

Et pour simplifier les notations **on note $P(Y = 1/X_1, X_2, X_3, \ldots.Xn) = p$**

Et c'est facile de montrer que $$logit(p) = log(\frac{p}{1-p}) = \sum_{j=1}^{n} \beta_j \ X_{ij}$$

Les régressions logistiques

le kit de l’utilisateur

- Savoir lire les coefficients de régression
- Ou la présentation des résultats sous forme d’odds-ratios
- - savoir repérer la qualité de la régression

Savoir lire la qualité de la régression =

→ **les seuils de significativité** comme pour les autres régressions.

→ **le pseudo R2** = pourcentage de l'inertie globale observée dans l'échantillon et pris en charge par le modèle = dont le modèle parvient à rendre compte. Ce pseudo-R2 est souvent faible.

→ **le pourcentage de concordance** doit être supérieur à 60% en principe. Il mesure la façon dont le modèle permet de prédire les probabilités individuelles correctement.

→ **Wald-R, la statistique de Wald** s'intéresse à la qualité de l'ensemble du modèle et plus à celle de chacun de ses paramètres. Mais le principe est le même, on teste la significativité statistique globale du modèle. On calcule une p-value associée à la cette statistique de wahl. Plus elle est faible, plus on a de chance que le modèle ne dise pas n'importe quoi.

Un exemple de ce qu’on obtient

Jérome Deauvieau dans le Bulletin de méthodologie statistique, 2010

On cherche à comprendre **la probabilité de connaître une mobilité professionnelle ascendante** :

- en partant d'un échantillon tiré de l'enquête FQP : des salariés du privé classés dans les professions intermédiaires
- On observe leur position professionnelle quelques années plus tard : certains sont devenus PCS3 et d'autres non
- On cherche donc à estimer la probabilité d'appartenir à la PC3 pour les individus classés quelques années plus tôt dans les professions intermédiaires.
- On cherche à estimer comment cette probabilité varie « toutes choses égales par ailleurs », en fonction de différentes caractéristiques des individus
  = leur âge, leur sexe, leur niveau de diplôme

Pour les variables à plus de deux modalités, on forme autant de variables dichotomiques qu'il y a de modalités dans la variable et on réalise ainsi un codage disjonctif complet. Par exemple, la variable âge, qui a trois modalités, sera transformée en trois variables dichotomiques âge 1 (code 1 si âge inférieur à 35 ans, sinon 0), âge 2 (code 1 si âge compris entre 35 et 45 ans, sinon 0), âge 3 (code 1 si âge supérieur à 45 ans, sinon 0).

Tableau 1. Le codage disjonctif complet

Variable	Modalité 1	Modalité 2	Modalité 3
1	1	0	0
2	0	1	0
3	0	0	1

On cherche une relation de la forme :

$$\ln \frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)} = \mathrm{B0} + \mathrm{B1diplome} + \mathrm{B2sexe} + \mathrm{B3age\ 2} + \mathrm{B4age\ 3}$$

Et on obtient avec les données de l'échantillon :

$$\ln \frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)} = -1,95 + 0,75 * \mathrm{diplôme} + 0,59 * \mathrm{sexe} - 0,29 * \mathrm{age\ 2} - 0,63 * \mathrm{age\ 3}$$

Présentation et lecture des résultats

Tableau 3. Expliquer le passage à la catégorie cadre

	Coefficient	
Constante	−1,95	Test
Sexe		
Femme	ref	
Homme	0,59	(p<0,01)
Diplôme		
Inférieur au bac	ref	
Supérieur au bac	0,75	(p<0,01)
Age		
Age 1	ref	
Age 2	−0,29	(p=0,22)
Age 3	−0,63	(p<0,01)

Source: INSEE, enquête FQP, 2003.

Ici pour on a codé

Pour une femme : Sexe = 0
Pour une homme : Sexe = 1

$$\ln \frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)} = -1,95 + 0,75 * \text{diplôme}$$
$$+ 0,59 * \text{sexe} - 0,29 * \text{age 2} - 0,63 * \text{age 3}$$

Tableau 3. Expliquer le passage à la catégorie cadre

	Coefficient	
Constante	−1,95	Test
Sexe		
Femme	ref	
Homme	0,59	(p<0,01)
Diplôme		
Inférieur au bac	ref	
Supérieur au bac	0,75	(p<0,01)
Age		
Age 1	ref	
Age 2	−0,29	(p=0,22)
Age 3	−0,63	(p<0,01)

Source: INSEE, enquête FQP, 2003.

Le logit des hommes est supérieur à celui des femmes, le logit des diplômés est supérieur à celui des non diplômés.
Pour la variable âge, on peut ordonner les : logit age 3 < logit age 2 < logit age 1.

Or logit P1 inférieur à logit de P2 implique que

$$\frac{P1}{1-P1} < \frac{P2}{1-P2}, \text{ et } \frac{1-P1}{P1} > \frac{1-P2}{P2}, \text{et } \frac{1}{P1} - 1 < \frac{1}{P2} - 1, \text{ et } \frac{1}{P1} > \frac{1}{P2}, \text{ enfin que P1} < \text{P2.}$$

Lecture du coefficient associé à une modalité

- Le signe : quand le coefficient est négatif, la modalité joue négativement sur la variable dépendante. C'est l'inverse quand le coefficient est positif.
- La valeur absolue : plus la variable joue un rôle important plus la valeur absolue du coefficient est élevée. A contrario plus il est proche de zéro plus l'influence de la caractéristique étudiée est faible.

Tableau 2. Calcul des logit du chaque situation

Situation	logit
Pas le bac, femme, age 1	−1,95
Pas le bac, femme, age 2	−2,24
Pas le bac, femme, age 3	−2,58
Pas le bac, homme, age 1	−1,36
Pas le bac, homme, age 2	−1,64
Pas le bac, homme, age 3	−1,99
Bac, femme, age 1	−1,20
Bac, femme, age 2	−1,48
Bac, femme, age 3	−1,83
Bac, homme, age 1	−0,60
Bac, homme, age 2	−0,89
Bac, homme, age 3	−1,23

Source: INSEE, enquête FQP, 2003.

$$\ln \frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)} = -1,95 + 0,75 * \text{diplôme} + 0,59 * \text{sexe} - 0,29 * \text{age 2} - 0,63 * \text{age 3}$$

on peut **en déduire des odds- ratio et souvent les résultats sont présentés directement en odds-ratios, souvent par rapport au groupe de référence**

Par exemple, pour comparer les probabilités des hommes et des femmes de devenir cadre, on peut remarquer que

$$\ln\frac{P1}{1-P1} - \ln\frac{P2}{1-P2} = 0,59,$$

Avec P1 = proba pour les hommes »toutes choes égales par ailleurs »

Et P2 = proba pour les femmes « toutes choses égales par ailleurs

Puis en déduire que :

$$\ln\frac{\dfrac{P1}{1-P1}}{\dfrac{P2}{1-P2}} = 0,59$$

Et enfin que :

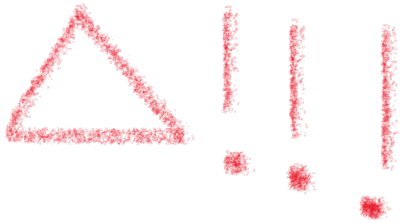
$$\exp\left(\ln\frac{\dfrac{P1}{1-P1}}{\dfrac{P2}{1-P2}}\right) = \exp(0,59) \qquad \text{donc} \quad \frac{\dfrac{P1}{1-P1}}{\dfrac{P2}{1-P2}} = 1,80.$$

Quelques réflexes utiles

Tableau 7. Régression logistique sur le vote d'extrême droite au 1er tour présidentiel 2002[a]

	B	Signif.	Exp (B)
Statut			
Indépendant	0,162	0,420	1,176
Salarié du privé	Référence		
Salarié du public	– 0,291	*0,041*	0,747
Chômeur	0,185	0,448	1,203
Classe			
0 attache ouvrière	– 0,071	0,614	0,932
1 attache	Référence		
2 attaches	0,016	0,930	1,017
Patrimoine			
0 ou 1 élément	0,015	0,917	1,015
2 éléments	Référence		
3 ou +	0,145	0,361	1,156
Revenu			
Moins de 10 000 francs mensuels	0,016	0,917	1,017
10 000-15 000 francs	Référence		
Plus de 15 000 francs	– 0,188	0,225	0,829
Sexe			
Homme	Référence		
Femme	– 0,415	*0,001*	0,660
Âge, diplôme			
Moins de 40 ans sans bac	0,800	*0,000*	0,2226
Moins de 40 ans avec bac	– 0,090	0,694	0,914
40 et + sans bac	0,566	*0,009*	1,762
40 et + avec bac	Référence		
Constante	– 2,058	*0,000*	0,128

a. La valeur absolue du coefficient B mesure le pouvoir prédictif des modalités de chaque variable une fois contrôlé l'effet des autres variables de la régression. Plus il est élevé plus la relation est forte. Le coefficient de la seconde colonne indique si la relation est significative sur le plan statistique. Plus il est faible, plus elle est significative.



	Modèle 1	Modèle 2	Modèle 3
R^2 de Cox et Snell	0,049	0,061	0,124
Sexe			
Homme	0,946	0,863*	0,961
Femme	*Réf.*	*Réf.*	*Réf.*
Âge			
18-24	0,430***	0,613**	1,037
25-34	0,763*	0,943	1,381**
35-49	0,662***	0,761**	1,064
50-64	0,605***	0,650***	0,764**
65 et plus	*Réf.*	*Réf.*	*Réf.*
Diplôme			
Sans diplôme	1,950***	2,391***	2,474***
BEPC	1,603***	1,952***	1,918***
Bac	1,302**	1,377**	1,360**
> bac	*Réf.*	*Réf.*	*Réf.*
Éléments de patrimoine			
0	0,470***	0,546***	0,713**
1	0,586***	0,635***	0,727**
2	0,784**	0,861	0,913
3-4	*Réf.*	*Réf.*	*Réf.*
Revenu par unité de consommation (transformation logarithmique)			
Q1 - Niveau de vie bas	0,637***	0,540***	*0,504*** (nl)*
Q2	0,657***	0,568***	0,504***
Q3	0,827	0,721**	0,628***
Q4	0,864	0,807	0,726**
Q5	0,739**	0,718**	0,667***
Q6 - Niveau de vie élevé	*Réf.*	*Réf.*	*Réf.*
Statut			
Indépendant, chef d'entreprise		2,571***	2,471***
Salarié du privé		1,352***	1,321***
Salarié du public		*Réf.*	*Réf.*
Religion			
Catholique pratiquant			6,705***
Catholique non pratiquant			2,912***
Autres religions			0,957
Sans religion			*Réf.*

Probabilité que les coefficients soient significatifs : *** p < 0,01, ** p < 0,05, * p < 0,1.
Note de lecture (nl) : dans le modèle 3, les chances d'avoir voté pour Nicolas Sarkozy sont divisées par presque deux (1/0,504 = 1,98) quand on a un bas niveau de vie plutôt que quand on a un niveau de vie élevé (Réf.).

Régression logistique du vote Nicolas Sarkozy au second tour de l'élection présidentielle en 2007.

Source RFSP, Mariette SINEAU et Viviane Le Hay, 2010
« Effet patrimoine, 30 ans après, le retour? »

Ligne 1 : l'odds-ratio augmente, quand on introduit de nouvelles variables, ce qui traduit …… ?
Globalement les seuils de significativité associés aux différentes modalités de la variables diminuent, ce qui traduit ….?

Les coefficients dans le tableau sont déjà sous forme d'odds-ratio par rapport au groupe de référence

Exp(B)	*A voté à toutes les élections*	*Appartient à 1 ou plusieurs associations*	*Parle politique en famille (souvent, de tps en tps)*	*« Le résultat de la présidentielle améliorera les choses en France »*	*Intérêt pour la politique (Beaucoup + assez)*	*« La démocratie fonctionne bien » (TB + assez)*	*A suivi la campagne tous les jours ou presque*
R^2 de Cox et Snell	0,049	0,045	0,047	0,038	0,083	0,055	0,060
Sexe							
Homme	1,046	1,405***	0,960	0,863**	1,654***	1,369***	1,348***
Femme	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Âge							
18-24	1,182	0,456***	1,171	0,586***	0,901	0,695**	0,552***
25-34	0,387***	0,446***	0,936	0,874	0,681***	0,584***	0,507***
35-49	0,434***	0,621***	0,902	0,783**	0,640***	0,738***	0,473***
50-64	0,549***	0,829*	0,989	0,706***	0,925	0,874	0,671***
65 et plus	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Diplôme							
Sans diplôme	0,827	0,402***	0,412***	1,897***	0,341***	0,485***	0,417***
BEPC	0,761***	0,506***	0,523***	1,740***	0,458***	0,601***	0,504***
Bac	0,970	0,667***	0,685**	1,190	0,631***	0,831	0,629***
> bac	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Éléments de patrimoine							
0	0,580***	0,506***	0,543***	0,455***	0,760**	0,546***	0,756**
1	0,760***	0,656***	0,661***	0,619***	0,832*	0,615***	0,641***
2	0,937	0,846*	0,838	0,749***	0,973	0,862	0,848
3-4	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Revenu par unité de consommation							
Q1	1,056	0,904	0,455***	0,716**	0,377***	0,531***	0,477***
Q2	0,831	0,865	0,522***	0,682***	0,454***	0,578***	0,629***
Q3	1,022	0,986	0,596***	0,755**	0,452***	0,670***	0,630***
Q4	0,977	1,021	0,544***	0,880	0,506***	0,862	0,659***
Q5	0,912	1,117	0,741**	0,889	0,681***	0,897	0,797*
Q6	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.	Réf.
Interaction revenus * patrimoine	Non Significatif[1]	NS	NS	NS	NS	NS	NS

	N'a confiance ni dans la gauche, ni dans la droite pour gouverner	*« Les responsables politiques ne se préoccupent pas de ce que pensent les gens »*	*Approuve l'occupation des bâtiments publics*	*A « peur en pensant à l'avenir »*
R^2 de Cox et Snell	0,051	0,055	0,057	0,069
Sexe				
Homme	0,840**	0,840**	1,053	0,500***
Femme	Réf.	Réf.	Réf.	Réf.
Âge				
18-24	1,721***	1,098	4,451***	1,851***
25-34	2,015***	1,732***	3,082***	1,907***
35-49	1,863***	1,672***	2,913***	1,532***
50-64	1,341***	1,629***	1,904***	1,344***
65 et plus	Réf.	Réf.	Réf.	Réf.
Diplôme				
Sans diplôme	1,106	2,018***	0,998	1,678***
BEPC	1,199*	2,001***	0,923	1,594***
Bac	1,083	1,714***	0,901	1,343**
> bac	Réf.	Réf.	Réf.	Réf.
Éléments de patrimoine				
0	1,760***	1,827***	1,568***	1,294**
1	1,489***	1,436***	1,571***	1,220*
2	1,261**	1,063	1,394***	1,197*
3-4	Réf.	Réf.	Réf.	Réf.
Revenu par unité de consommation				
Q1	2,346***	1,930***	1,253*	2,158***
Q2	2,168***	1,888***	1,500***	2,206***
Q3	1,974***	2,113***	1,142	1,775***
Q4	1,666***	1,850***	1,218*	1,774***
Q5	1,554***	1,387***	1,177	1,459***
Q6	Réf.	Réf.	Réf.	Réf.
Interaction revenus * patrimoine	NS	NS	NS	NS

	Appartenance politique : droite	*Auto-positionnement à droite*	*Préférence pour gouverner : droite*	*Proximité partisane à droite (UDF + UMP)*	*Attributs de droite : 1 ou plusieurs*
R^2 de Cox et Snell	0,051	0,047	0,054	0,064	0,046
Sexe					
Homme	1,088	1,061	1,069	1,111	1,028
Femme	Réf.	Réf.	Réf.	Réf.	Réf.
Âge					
18-24	0,432***	0,483***	0,476***	0,440***	0,560***
25-34	0,693***	0,754**	0,604***	0,602***	0,880
35-49	0,578***	0,587***	0,593***	0,547***	0,727***
50-64	0,638***	0,711***	0,690***	0,618***	0,833*
65 et plus	Réf.	Réf.	Réf.	Réf.	Réf.
Diplôme					
Sans diplôme	1,221	1,285*	1,241	0,703***	0,831
BEPC	1,195	1,283**	1,183	0,836*	0,899
Bac	1,055	1,099	1,023	0,847	0,755**
> bac	Réf.	Réf.	Réf.	Réf.	Réf.
Éléments de patrimoine					
0	0,425***	0,403***	0,420***	0,469***	0,497***
1	0,489***	0,465***	0,522***	0,557***	0,498***
2	0,749***	0,702***	0,682***	0,853	0,754***
3-4	Réf.	Réf.	Réf.	Réf.	Réf.
Revenu par unité de consommation					
Q1	0,522***	0,535***	0,410***	0,408***	0,519***
Q2	0,625***	0,591***	0,496***	0,486***	0,603***
Q3	0,746**	0,637***	0,570***	0,707***	0,808
Q4	0,823	0,782**	0,706***	0,746**	0,725***
Q5	0,808*	0,729***	0,678***	0,796**	0,801*
Q6	Réf.	Réf.	Réf.	Réf.	Réf.
Interaction revenus * patrimoine	NS	NS	NS	NS	NS

Un exemple de régressions logistiques emboîtées

Référence = un travail de Claude THELOT avec Jean-Paul CAILLE, un travail non publié, présenté aux 39èmes journées de la statistiques, en 2007.

Les enfants étrangers réussissent moins bien à l'école?

- Pannel = les élèves entrés en 6ème en septembre 1989 et né le 5 d'un mois
- Taille de l'échantillon = 17314
- Variable dépendante = obtenir le bac après 7 ans dans le secondaire (= sans avoir redoublé)
- Variable d'intérêt = nationalité de l'enfant
- 30.9% des élèves de l'échantillon ont obtenu leur bac 7 ans près être entrés en 6ème
- 7.9% des élèves sont étrangers

Les résultats du tri croisé

31.8% des entrants en 6ème de nationalité française ont obtenu leur bac en 7 ans
C'est le cas de 19.6% des élèves étrangers.

On peut en déduire un odds ratio = 0.562 = le fait d'être étranger divise en apparence par deux la chance relative d'avoir son bac en 7 ans plutôt que de ne pas l'avoir.

Ce lien statistique entre la variable d'intérêt et la variable dépendante permet-il de conclure à un lien de cause à effet entre la nationalité et la réussite scolaire mesurée par la réussite au bac sans avoir jamais redoublé?

Les autres variables qui peuvent influencer la réussite scolaire

- Ressources socio-économiques PCS du chef de famille, statut d'activité de la mère, nombre moyen de personne par pièce comme « proxy » du revenu (modèle II)
- Ressources culturelles diplôme le plus élevé de la mère, le fait que l'un des parents a (ou n'a pas) suivi une formation postscolaire à son initiative, le fait que l'un des parents est (ou n'est pas) enseignant, le fait que l'enfant a (ou n'a pas) un frère ou une soeur plus âgé(e) scolarisé(e) dans un lycée ou dans l'enseignement supérieur
- Autres aspects objectifs de la situation familiale structure de la famille, nombre d'enfants, rang de naissance de l'enfant

Les autres variables qui peuvent influencer la réussite

- Mesure initiale des performances scolaires test d'entrée en sixième = une variable ordinale qui correspond à une distribution en quartiles.
- Une analyse préliminaire montre que les enfants étrangers obtiennent des scores légèrement plus faibles en moyenne à ces tests. (modèle IV)
- Aspirations scolaires de la famille le fait que les parents souhaitent ou non que leurs enfants poursuivent leurs études au-delà de 20 ans, le fait qu'ils déclarent ou pas qu'un diplôme du supérieur est utile pour trouver un emploi.
- Une analyse préliminaire montre que « toutes choses égales par ailleurs », les familles étrangères manifestent un degré plus élevé d'aspiration pour leurs enfants. (modèle V)

Les résultats des régressions logistiques emboîtées

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
Élèves français	ref	ref	ref	ref	ref
Élèves étrangers (coeffcient de la régression)	-0.666***	-0.10	+0.25**	+0.42***	+0.32**

Note : ici les coefficients n'ont pas été transformés en odds ratios

La décomposition des effets

1. Réussite dans le parcours scolaire = admission en terminale 6 ans après l'entrée en sixième

	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
étranger	-0.56***	-0.00	+0.35***	+0.57***	+0.45***

2. Réussite à l'examen lui-même (réussir le bac quand on est entré en terminale 6 ans après être entré en 6ème)

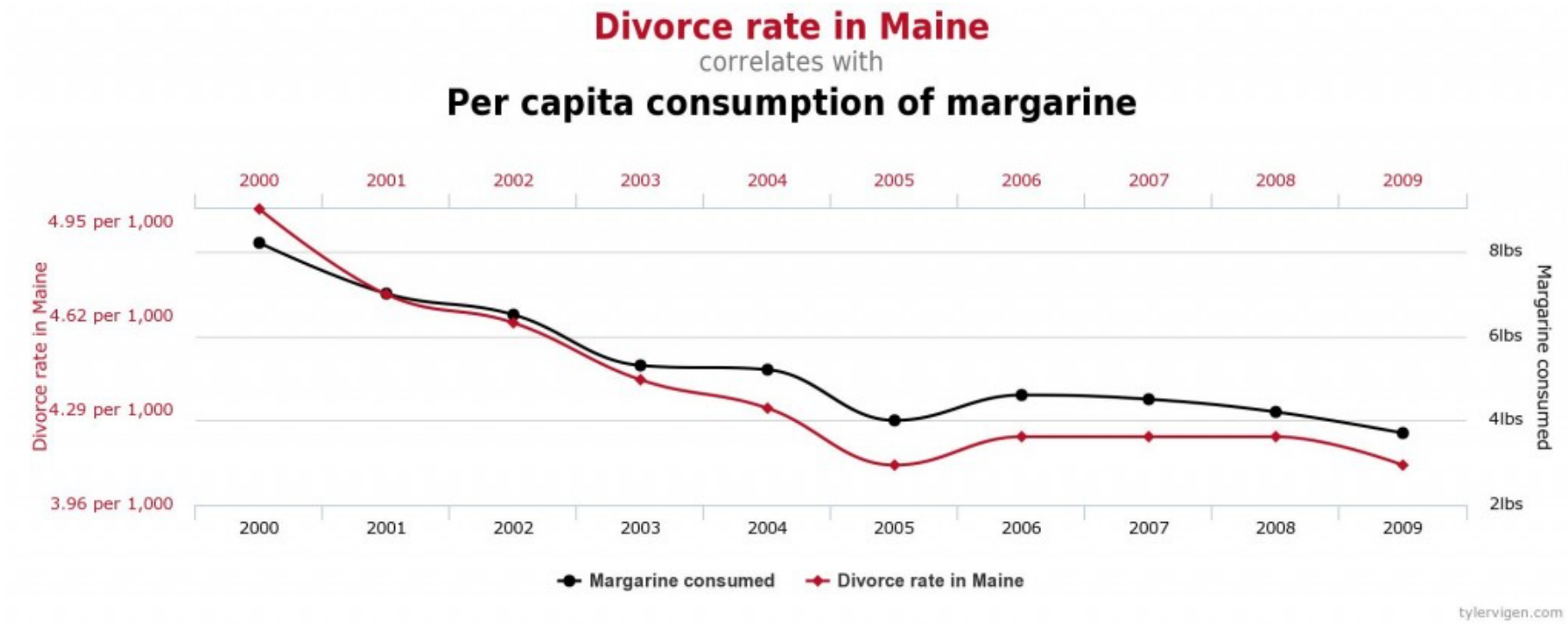
	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
étranger	-0.58***	-0.37*	+0.30*	-0.23	-0.23

Et dernier petit rappel pour éviter une « logical fallacy » fréquente dans l’analyse de données

L’effet cigogne ou « cum hoc ergo propter hoc »

prétendre que si deux variables sont corrélées alors l’une est nécessairement la cause de l’autre, c’est se livrer à une forme de sophisme, un exemple de « logical fallacy ».

La margarine permet d'éviter les divorces ?



Corrélation et causalité

Lorsqu'on observe une corrélation entre deux phénomènes A et B, il existe au moins six façons d'en rendre compte :

- A est la cause de B
- B est la cause de A
- A et B ont pour cause un même phénomène-source C, qu'on appelle la variable cachée.
- A est la cause de C qui est la cause de B (ou l'inverse)
- A est la cause de B et dans le même temps B est la cause de A (les deux phénomènes se renforcent)
- La co-occurrence de A et B est une pure coïncidence!

Partie 2. Raisonner toutes choses inégales par ailleurs

1. Les mises en garde contre le raisonnement toutes choses égales par ailleurs en sciences sociales
2. Les analyses factorielles au service d'une analyse relationnelle de l'espace social

1. Maurice HALBWACHS, dans « la statistique en sociologie », 1935

« combien de temps vivraient les Français si, restant français, ils vivaient dans les mêmes conditions physiques et sociales que les Suédois? Combien de temps vivraient les Allemands si, restant allemands, ils vivaient dans les mêmes conditions que les Français?

« Cela revient, comme l'observait Simiand à propos d'une comparaison économique récente entre les niveaux de vie des différents pays, à se demander **comment vivrait un chameau, si, en restant chameau, il était transporté dans les régions polaires**, et comment vivrait un renne si, en restant renne, il était transporté dans le Sahara. »

« Comme si l'on avait affaire à des hommes qui ne naissent, ne se marient, ne meurent dans aucune région définie de quelque manière, quant aux coutumes familiales, religieuses, juridiques, économiques. Mais de même que l'homo oeconomicus, un tel homo demographicus est une abstraction trop détachée de la réalité pour nous apprendre quoi que ce soit de réel. »

« on risque de **superposer aux groupes réels des groupes fictifs** qui ne paraissent correspondre aux premiers que parce qu'ils ne sont que ces premiers en effet, mais privés d'une grande partie de leur contenu. Est-on bien sûr alors que ce qu'on a ainsi écarté pour des motifs de simplification n'était pas l'essentiel, ce sans quoi on ne peut expliquer la réalité? »

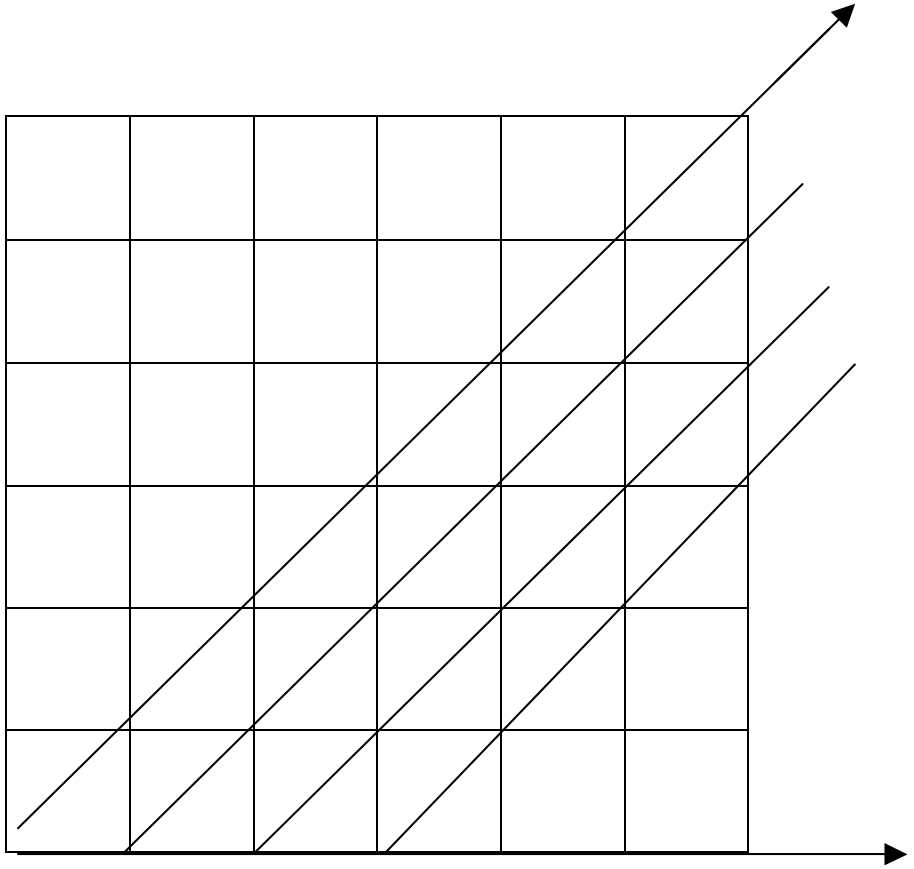
2. Passeron : « Ce que dit un tableau et ce qu'on en dit » dans *Le raisonnement sociologique* . Un espace non poppérien du raisonnement naturel, 1991

	Très fréquente Y_1	Fréquence intermédiaire Y_2	Faible fréquence Y_3
Jeune X_1	80%	15%	5%
Age intermédiaire X_2	20%	60%	20%
Vieux X_3	5%	15%	80%

Raisonner à âge comparable?

Effet d'âge, effet de génération ou effet de période :
trois types d'enquêtes pour essayer de les isoler ou le modèle à trois tem

le diagramme de Lexis (1880), facilite la mise en relation de l'optique longitudinale et transversale.



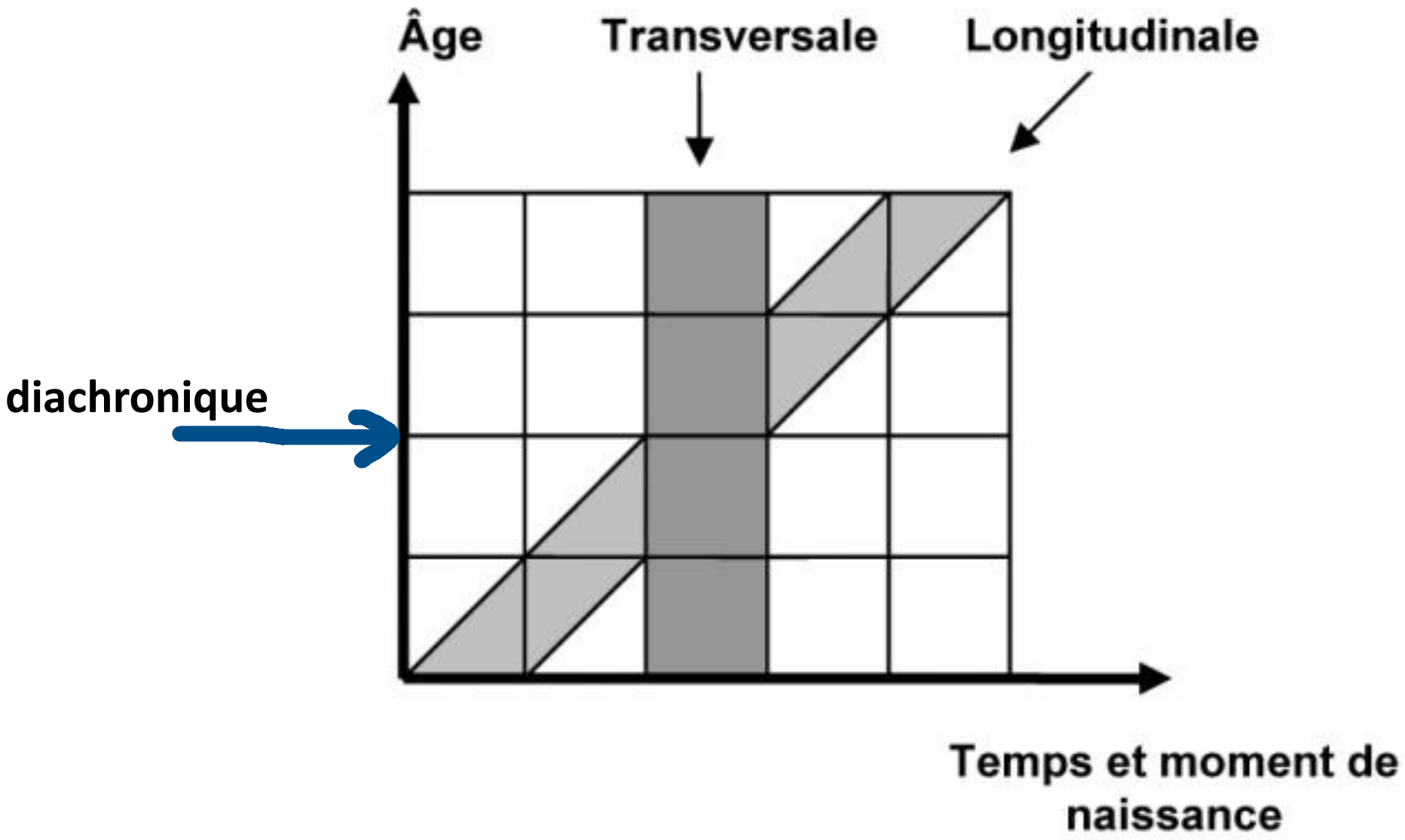
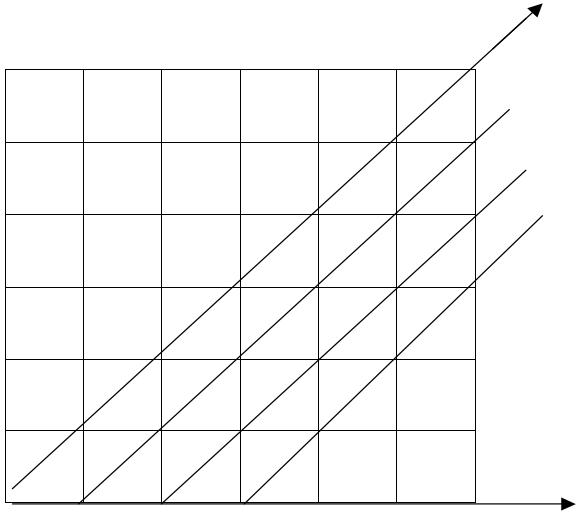


Diagramme de Lexis – Analyses transversale et longitudinale

Effet d'âge, effet de génération ou effet de période :
trois types d'enquêtes pour essayer de les isoler ou le modèle à trois tem



- Analyse **transversale** : on collecte des données pour différentes classes d'âge à une même date, on raisonne à période donnée, mais on risque de confondre effet d'âge et effet de génération
- on peut suivre des cohortes dans des études **longitudinales**, on peut raisonner alors à cohorte donnée et comparer l'évolution de la pratique étudiée au sein des cohortes à différents moments du cycle de vie, mais par définition elles n'ont pas le même âge à la même période.
- Si on raisonne sur une même classe d'âge à des périodes différentes, dans une approche **diachronique** de la classe d'âge, on risque de confondre effets de génération et effet de période. Exemple avoir 20 ans en 1968, en 1986, en 2002, en 2008

3. Marie Duru-Bellat
Les inégalités sociales à l'école. Genèse et mythes, 2002

« Si la **modélisation** ouvre au chercheur en sciences sociales des horizons heuristiques, il reste que la clause « toutes choses égales par ailleurs » sur laquelle elle se fonde présente **un risque de «sociologie fiction»** redoutable (discuté notamment par Passeron, 1991). L'estimation de modèles multivariés est **une fiction de raisonnement expérimental**, souvent «limite», précisément parce que le raisonnement expérimental sur lequel ils reposent est évidemment très éloigné de la réalité.

Si on admet sans peine qu'on peut introduire, pour expliquer les choix d'orientation, à la fois les notes scolaires et le sexe (variables corrélées), pour évaluer un effet du sexe « toutes choses égales par ailleurs » (effet net restant lui-même à expliquer), le sociologue sera plus gêné devant l'introduction simultanée de l'origine sociale et de l'origine ethnique, si on entend en déduire un effet de l'origine ethnique « toutes choses égales par ailleurs ». **La quête de l'effet pur tourne ici à la sociologie fiction** : dans la réalité, la distribution des niveaux d'instruction des parents (de même que la plupart de leurs caractéristiques sociales) est tout sauf égale, entre enfants français et étrangers. Le sens même de cette variable est défini dans son articulation avec d'autres » (Marie Duru-Bellat, 2002, pp. 48-49).

4. Emmanuel PIERRU et Alexis SPIRE « le crépuscule des catégories socioprofessionnelles », RFSP, 2008

« l'inclusion dans un même modèle statistique de la catégorie socioprofessionnelle avec des variables telles que le diplôme ou encore le statut d'activité a pour effet – selon l'heureuse expression employée par François Héran – **d'« essorer » la catégorie sociale en lui retirant – par décomposition des « effets purs » – les propriétés synthétiques qui la rendent précisément agissante**. Qu'elle soit enregistrée à un niveau agrégé ou à un niveau plus fin, la PCS synthétise en définitive un faisceau de propriétés qui convient mal au raisonnement économétrique. »

Jérôme Deauvieau se réfère aussi à François « Héran note ainsi sur le raisonnement en termes de régression multiple à propos des scolarités des enfants d'origine étrangère que la réussite aux évaluations scolaires ne diffèrent pas de celle des autres élèves (toutes choses égales par ailleurs),

mais ajoute immédiatement **qu'il reste que les enfants d'origine étrangère accueillis dans les collèges et les lycées ne se présentent jamais « toutes choses égales par ailleurs « , mais bien, si l'on peut dire, « toutes choses inégales réunies »;** (Héran, 1996). »

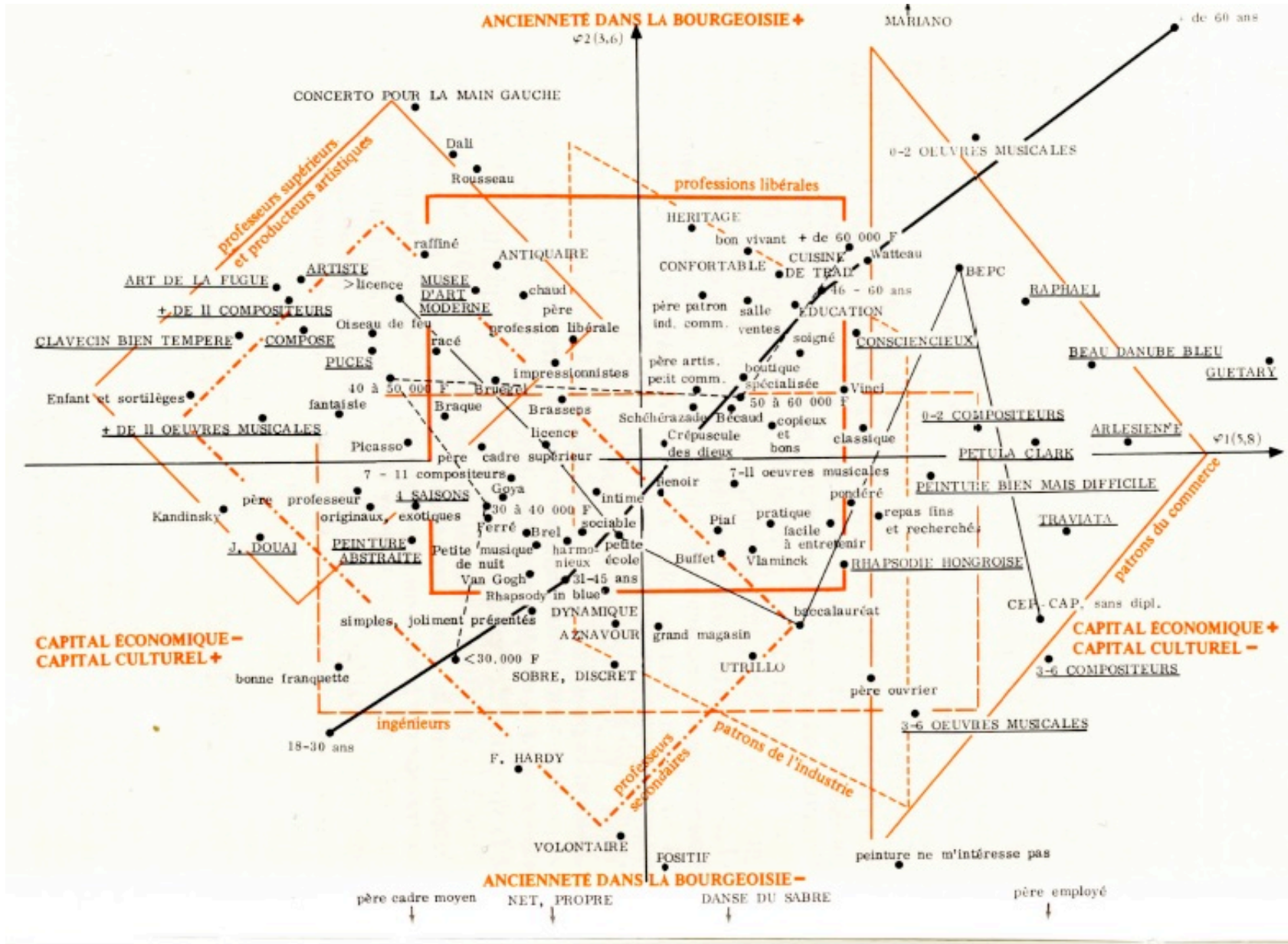
Et Bourdieu?

« **j’utilise beaucoup l’analyse des correspondances**, parce que je pense que c’est une procédure relationnelle dont la philosophie exprime pleinement ce qui selon moi constitue la réalité sociale. C’est une procédure qui « pense » relationnellement, et c’est ce que j’essaie de faire avec le concept de champ »

(préface à l’édition allemande de *Le métier de sociologue,* coécrit avec CHAMBOREDON et PASSERON, 1968 pour l’édition française*)*

« ce qu'il faut dénoncer ce n'est pas l'usage de la statistique mais le fétichisme de la statistique(…)
la magie terrifiante et fascinante des chiffres, outre qu'elle tient le lecteur novice à distance respectueuse, permet au spécialiste de faire l'économie de sa réflexion et de son intelligence et de s'en remettre à ce que Leibniz appelait *l'évidence aveugle* du calcul et des symboles.»
Travail et travailleurs en Algérie, avec Alain DARBEL, en 1963

« la sociologie serait moins vulnérable à la tentation de l'empirisme s'il suffisait de lui rappeler avec Poincaré que *les faits ne parlent pas* »,
Le métier de sociologue, avec Jean-Claude PASSERON 1968


ANCIENNETE DANS LA BOURGEOISIE +
φ2 (3,6)
ANCIENNETE DANS LA BOURGEOISIE −
φ1(5,8)
CAPITAL ECONOMIQUE −
CAPITAL CULTUREL +
CAPITAL ECONOMIQUE +
CAPITAL CULTUREL −
professeurs supérieurs et producteurs artistiques
professions libérales
ingénieurs
professeurs secondaires
patrons de l'industrie
patrons du commerce
CONCERTO POUR LA MAIN GAUCHE
Dali
Rousseau
ART DE LA FUGUE
ARTISTE
>licence
MUSEE D'ART MODERNE
+ DE 11 COMPOSITEURS
CLAVECIN BIEN TEMPERE
COMPOSE
Oiseau de feu
PUCES
Enfant et sortilèges
+ DE 11 OEUVRES MUSICALES
Picasso
Kandinsky
J. DOUAI
PEINTURE ABSTRAITE
père professeur
originaux, exotiques
4 SAISONS
père cadre supérieur
7 - 11 compositeurs
Goya
Ferré
Brel
Petite musique de nuit
Van Gogh
Rhapsody in blue
DYNAMIQUE
simples, joliment présentés
bonne franquette
<30.000 F
SOBRE, DISCRET
18-30 ans
F. HARDY
VOLONTAIRE
père cadre moyen
NET, PROPRE
raffiné
ANTIQUAIRE
chaud
père profession libérale
racé
impressionnistes
40 à 50 000 F
Bruckner
fantaisie
Braque
Brasseps
licence
intime
sociable
harmonieux
petite école
31-45 ans
AZNAVOUR
grand magasin
HERITAGE
bon vivant + de 60 000 F
CONFORTABLE
CUISINE DE TRAIT
père patron ind. comm.
salle ventes
46 - 60 ans
EDUCATION
soigné
père artis. petit comm.
boutique spécialisée
50 à 60 000 F
Schéhérazade
Bécaud
copieux et bons
Crépuscule des dieux
Renoir
Piaf
Buffet
Vlaminck
pratique
facile à entretenir
pondéré
repas fins et recherchés
7-11 oeuvres musicales
UTRILLO
père ouvrier
FOSITIF
DANSE DU SABRE
peinture ne m'intéresse pas
père employé
MARIANO
+ de 60 ans
0-2 OEUVRES MUSICALES
BEPC
RAPHAEL
CONSCIENCIEUX
BEAU DANUBE BLEU
GUETARY
Vinci
classique
0-2 COMPOSITEURS
PETULA CLARK
ARLESIENNE
PEINTURE BIEN MAIS DIFFICILE
TRAVIATA
RHAPSODIE HONGROISE
baccalauréat
CEP-CAP, sans dipl.
3-6 COMPOSITEURS
3-6 OEUVRES MUSICALES

Pour aller plus loin sur les affinités entre Bourdieu et les analyses des correspondances

Un article de Frédéric LEBARON :

« L'analyse géométrique des données dans un programme de recherche sociologique : Le cas de la sociologie de Bourdieu » dans la revue MODULAD, numéro 42, 2010

Quelques extraits de l'article de Lebaron

- Une critique par Bourdieu et les statisticiens qui travaillent avec lui comme DARBEL des modèles de régression : « ils développent une conception plus « structurale » de la causalité en sciences sociales que celle de la sociologie des variables, de la démographie ou de l'économie quantitative qui se développent alors : il s'agit d'étudier les effets globaux **d'une structure complexe d'interrelations, qui sont irréductibles à la combinaison des « effets purs » de variables indépendantes.** »

Lebaron (suite)

- « A la fin des années 1960, Bourdieu se tourne vers l'analyse des données, dont il perçoit l'affinité élective avec sa propre théorie structurale du monde social. Il intègre alors l'idée que si la quantification doit se développer en sciences sociales, elle doit être **multidimensionnelle**. Elle doit permettre, dans un premier temps, d'opérationnaliser **les différentes dimensions fondamentales de l'espace social, c'est-à-dire les différentes espèces de capitaux, économique, culturel, social et symbolique**; l'étape suivante consistant à **les combiner afin de fournir une modélisation géométrique des données**. »

Lebaron (suite)

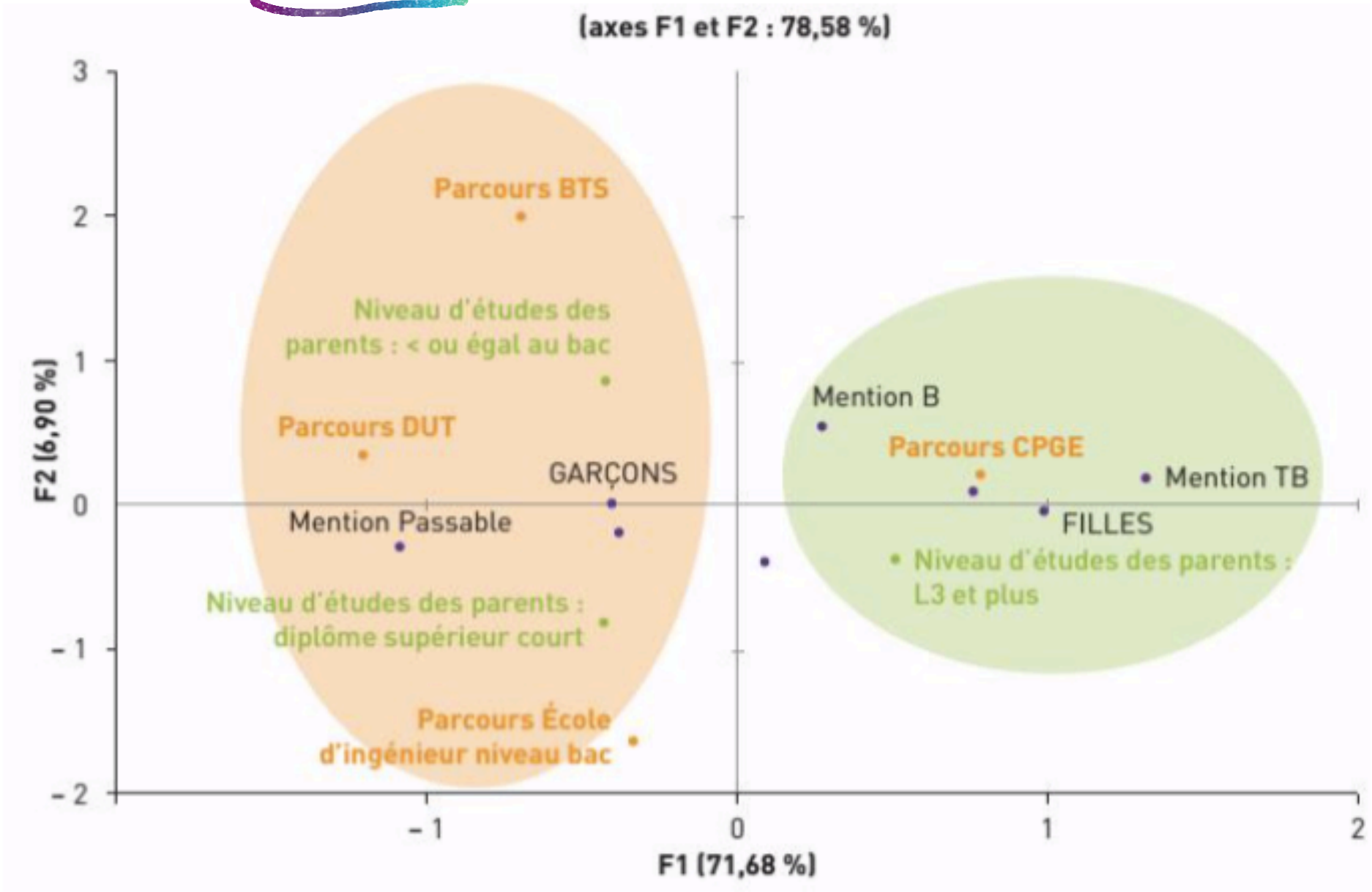
- Première application de l'ACM dans « l'anatomie sociale du goût », 1976. Repris ensuite dans La distinction.
- Les goûts apparaissent comme des variables actives
- Les indications sur la position dans l'espace social sont traitées comme variables supplémentaires
- Résultat graphique : « cette technique de visualisation donne une intuition des relations sociologiques entre l'espace des goûts (styles de vie) et l'espace des positions sociales » : on **visualise une homologie entre ces deux dimensions**
- Ensuite Bourdieu est resté très fidèle à cette méthodologie

Lire une analyse factorielle

Une méthode au service d’une représentation multidimensionnelle de l’espace social

Un exemple
choisi dans un sujet d’oral
(Lyon 2018)

Les ingénieurs un parcours différencié selon le genre



Source des données :

Enquête Génération 2010 du CEREQ. Elle fournit des informations sur le parcours scolaire et sur l'insertion professionnelle de 33 000 individus. Afin de mettre en évidence les différenciations du parcours selon le genre pour un même diplôme final, l'analyse est restreinte aux individus ayant obtenu un diplôme d'ingénieur.

Lecture :

Analyse des correspondances multiples (ACM), pour analyser le parcours d'études des ingénieurs en fonction du sexe, du capital humain et du capital social. L'axe vertical oppose à droite, les parcours des filles associés à des niveaux de capital humain et social élevés, et à gauche, les parcours des garçons, associés à des niveaux moindre de capital humain et social.

Source du document :

Jaoul-Grammare, 2018

Cette analyse met en évidence la sélection tant scolaire que sociale des parcours d'ingénieurs, mais souligne également que pour parvenir à un diplôme d'ingénieur, les filles doivent passer par la « voie royale » tant sur le plan du capital humain (mention B ou TB et CPGE, Bac S, « en avance »[20]) que social (parents ayant un niveau d'études supérieur à L3, cadres[21]). Nous rejoignons ici de précédents résultats [Jaoul-Grammare et Maddalou, 2013] selon lesquels entre 1992 et 2004, les inégalités injustes – c'est-à-dire les inégalités ne résultant pas d'un choix personnel de l'individu comme les inégalités liées au genre –, n'ont pas baissé dans l'enseignement supérieur français.

Objectifs et principes de l’analyse

- Il est difficile de projeter un espace à n dimensions sur l’espace à deux dimensions de la feuille.
- On commence par réduire le nombre de dimensions de l’espace en construisant des facteurs à partir des variables. **Les facteurs** sont des synthèses de différentes modalités de réponse, des méta-variables.

Résumés des principes d'une analyse factorielle

- L'analyse des correspondances cherche à synthétiser l'information contenue dans le tableau de contingence, en repérant des **facteurs,** qui condensent une partie des variations observées dans les réponses de l'échantillon. On dit que ces facteurs contribuent à une part de l'inertie totale.
- **L' inertie totale** mesure l'ampleur des écarts entre les réponses qu'on observe et les réponses qu'on observerait si les variables n'étaient pas corrélées entre elles, si elles étaient parfaitement indépendantes.
- On peut mesurer la part de cette inertie totale expliquée par chacun des facteurs
- On peut aussi pour chaque facteur mesurer la contribution de chaque variable active à sa construction, ce qui permet ensuite un commentaire sociologique des graphiques qui figurent ces facteurs.
- On ajoute ensuite des **variables supplémentaires** liées ici à l'appartenance sociale des individus : niveau de diplôme, niveau de revenu, PCS, âge et sexe. Elles ne participent pas à la construction des facteurs, mais elles permettent d'interpréter les résultats en les reliant à l'appartenance sociale des individus interrogés.

Exemple des relations entre deux variables : le sexe et les genres musicaux préférés

On construit un tableau des effectifs

	Rock	Jazz	Opéra	Chansons françaises	total
femme	15	9	2	24	50
homme	0	5	15	30	50
total	15	14	17	54	100

Exemple des relations entre deux variables : le sexe et les genres musicaux préférés

On le transforme en tableau

des fréquences observées

	Rock	Jazz	Opéra	Chansons françaises	total
femme	15%	9%	2%	24%	50%
homme	0%	5%	15%	30%	50%
total	15%	14%	17%	54%	100%

Exemple des relations entre deux variables : le sexe et les genres musicaux préférés

On construit un tableau des fréquences théoriques

	Rock	Jazz	Opéra	Chansons françaises	total
femme					50%
homme					50%
total	15%	14%	17%	54%	100%

Exemple des relations entre deux variables : le sexe et les genres musicaux préférés

On construit un tableau des fréquences théoriques (celles qu'on observeraient si les variables étaient indépendantes)

	Rock	Jazz	Opéra	Chansons françaises	total
femme	7.5%	7%	8.5%	27%	50%
homme	7.5%	7%	8.5%	27%	50%
total	15%	14%	17%	54%	100%

Exemple des relations entre deux variables : le sexe et les genres musicaux préférés

On construit un tableau des écarts à l'indépendance (écarts entre fréquences observées et fréquences théoriques)

	Rock	Jazz	Opéra	Chansons françaises	total
femme	7.5%	2%	-6.5%	-3%	
homme	- 7.5%	-2%	6.5%	3%	
total					

Exemple des relations entre deux variables : le sexe et les genres musicaux préférés

On construit un tableau des contributions au khi-deux

Khi-deux = (écart à l'indépendance)2/fréquence théorique

	Rock	Jazz	Opéra	Chansons françaises	total
femme	7.5	0.57	4.97	0.33	
homme	7.5	0.57	4.97	0.33	
total					26.74

Exemple des relations entre deux variables : le sexe et les genres musicaux préférés

On construit un tableau des contributions au khi-deux en valeur relative

	Rock	Jazz	Opéra	Chansons françaises	total
femme	28.05%	2.13%	18.59%	1.23%	50%
homme	28.05%	2.13%	18.59%	1.23%	50%
total	56.1%	4.26%	37.17%	2.46%	100%

Lecture : 56% de l'inertie totale est due à la modalité Rock et 93.25% de l'inertie totale est expliquée par la combinaison des modalités Rock et Opéra. Un facteur qui combinerait ces deux modalités épuiserait à lui tout seul plus de 90% de l'inertie totale observée dans l'échantillon.

Présentation d’un premier exemple d’ analyse des correspondances multiples

- Philippe COULANGEON, « La stratification sociales des goûts musicaux », Revue française de sociologie, 2003
- Les données sont issues de l’enquête de 1997 du ministère de la culture sur les pratiques culturelles
- Il travaille sur la question : « quels sont les genres musicaux que vous écoutez le plus souvent? »

Objectifs de l’analyse

- Tester deux modèles d’interprétation des pratiques culturelles sans les opposer :
- omnivore/univore de Peterson (1996)
- Celui de la légitimité culturelle initiée par Bourdieu dans *La distinction* (1979)
- Et pour remplir cet objectif : mener une analyse des correspondances multiples (ACM) pour « faire apparaître la variété des combinaisons entre les différents genres écoutés le plus souvent et le degré d’éclectisme des pratiques. »

Ici Coulangeon cherche à étudier les relations entre les réponses fournies par plus de 4000 individus.

Il y a 17 **variables actives** qui correspondent chacune à un type de réponse possible et une 18ème variable active qui correspond au nombre de genres cités dans la réponse, qui est un indicateur d'éclectisme.

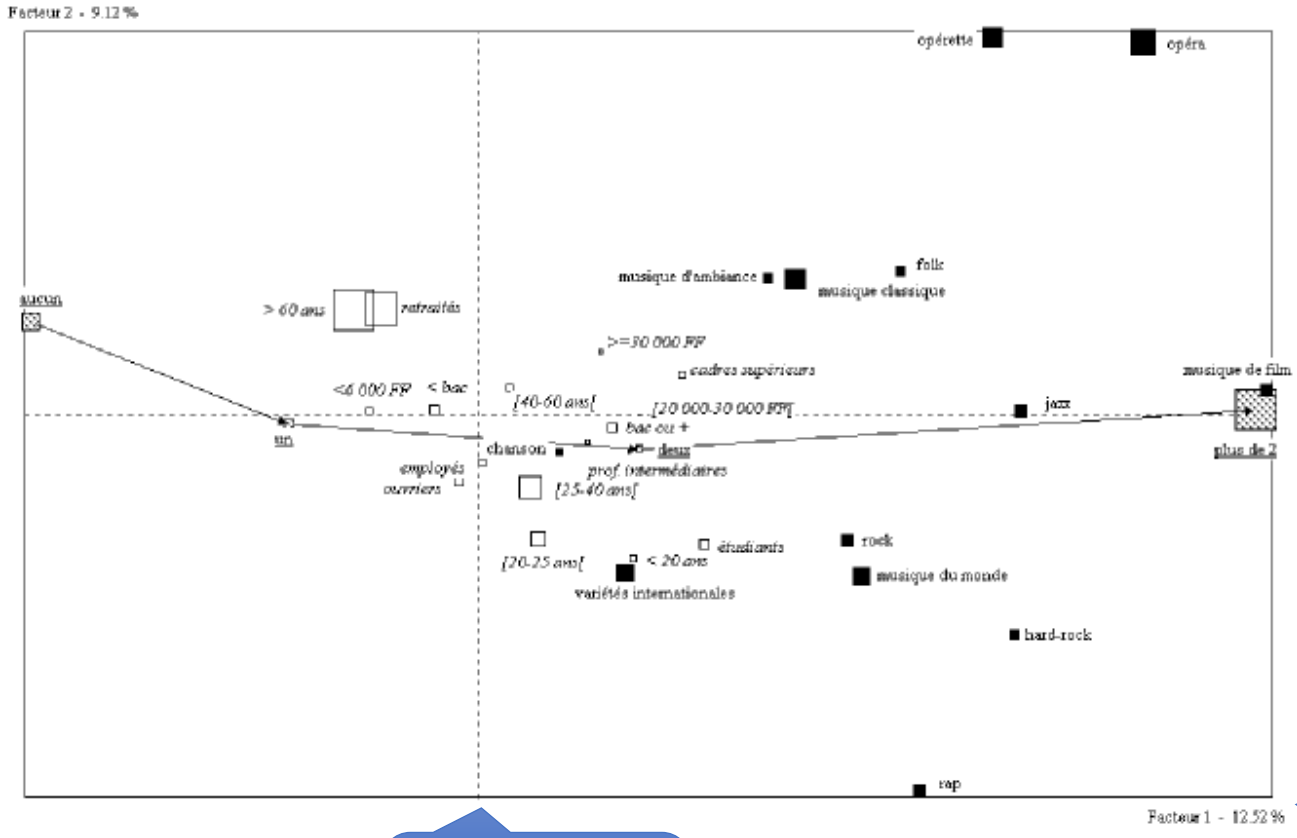
Il obtient **un vaste tableau de 4000 lignes**, où pour chaque individu sont répertoriées le ou les genres qu'il a cités et le nombre total de genres qu'il a cités, très long à lire.

C'est ce tableau qu'on cherche à **rendre lisible grâce à la synthèse des informations qui y sont contenues**

Dans les articles, c'est souvent **la représentation graphique des résultats de l'ACM** qui nous est fournie

Représentation graphique des résultats et interprétation

FIGURE I. – *L'espace des goûts musicaux (I). Plan des deux premiers facteurs de l'ACM*



Inertie du facteur 2 : 9.12%

Axe horizontal : facteur 1

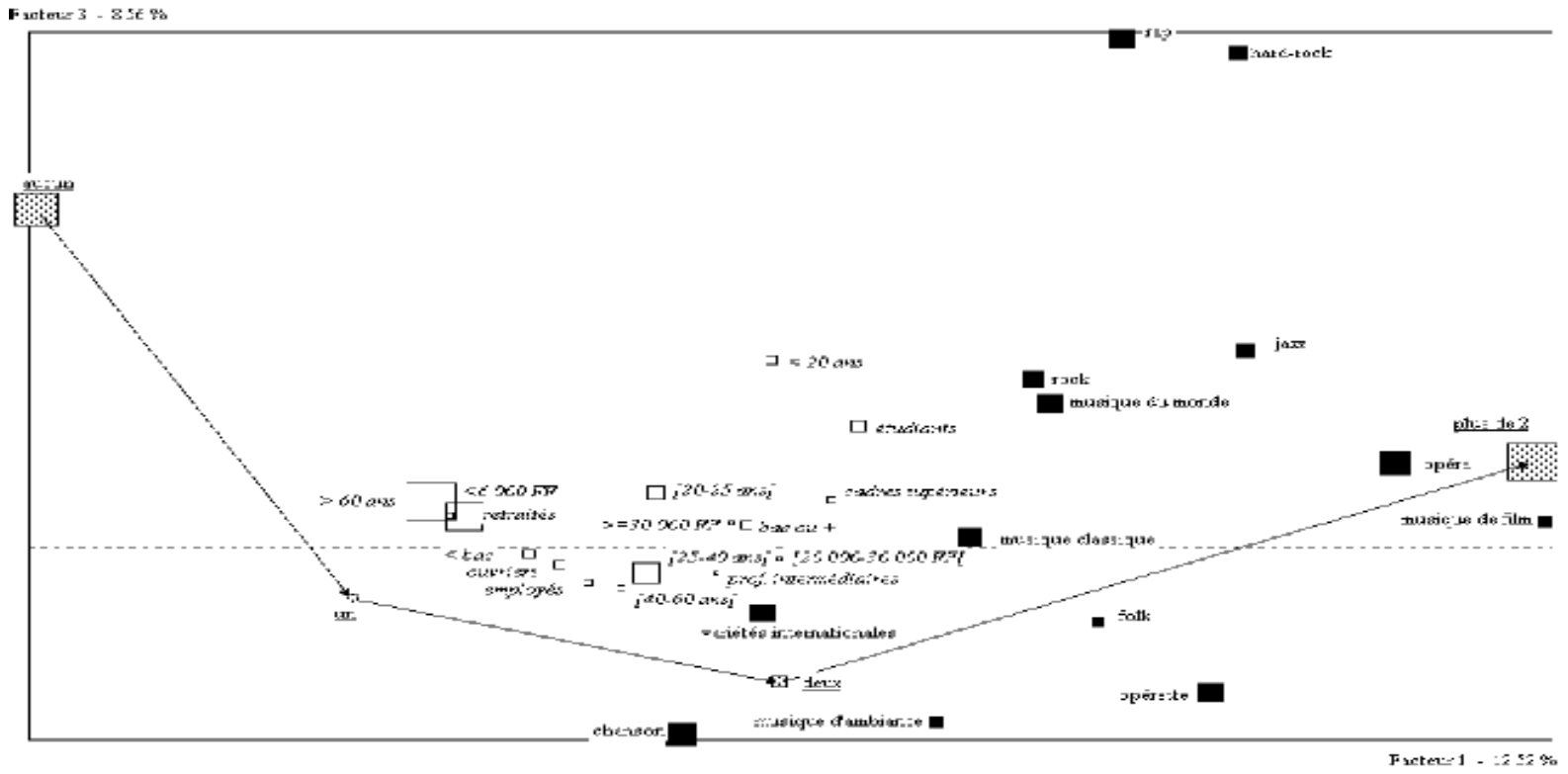
Inertie du facteur 1 : 12.5%

Axe vertical : facteur 2

Représentation graphiques des résultats et interprétation

Inertie facteur 3 : 8.16%

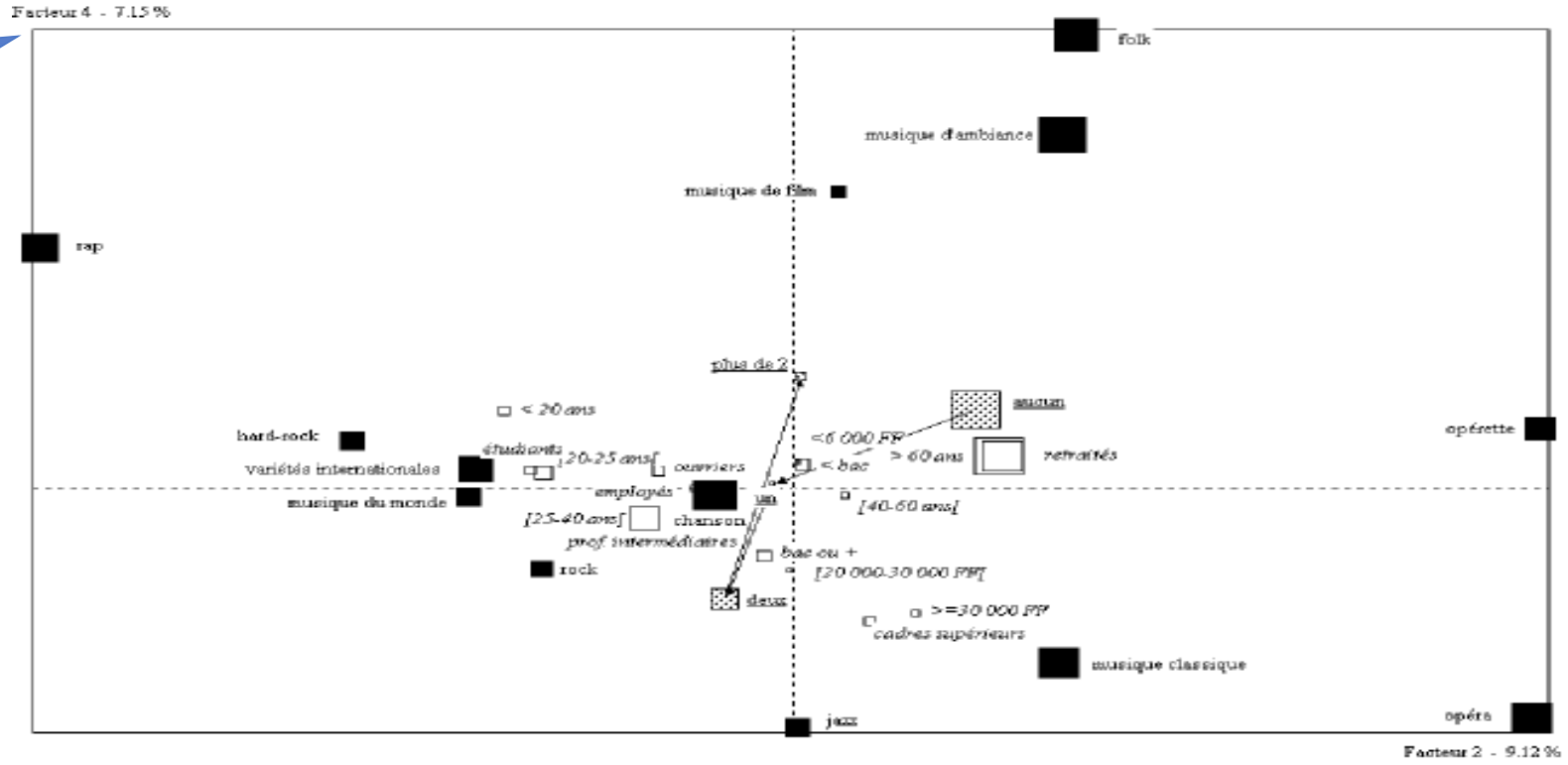
FIGURE II. – *L'espace des goûts musicaux (II). Plan des facteurs 1 et 3*



Représentation graphique des résultats et interprétation

Inertie facteur 4 : 7.15%

FIGURE III. – *L'espace des goûts musicaux (III). Plan des facteurs 2 et 4*



Une typologie des goûts musicaux construite à partir de cette ACM

- Classe 1 : trois genres qui relèvent de la musique savante, dont la musique classique, l'opéra et le jazz (20% de l'échantillon)
- Classe 2 : des musiques à usages variés (ambiance, musique de danse, musique de film,…) et un goût pour l'opérette (13%)
- Classe 3 : rap, rock, hard rock, musiques du monde, variétés internationales (8%)
- Classe 4 : un seul genre cité : les variétés internationales (44%)
- Classe 5 : aucun genre cité (14%)

La répartition des différentes PCS dans les classes de la typologie

TABLEAU III. – ***Distribution des catégories socioprofessionnelles par classes*** (en pourcentage)

	Classe I	Classe II	Classe III	Classe IV	Classe V	Total
Agriculteurs	10	16	4	47	23	100
Patrons de l'industrie et du commerce	20	10	12	45	13	100
Cadres supérieurs	54	9	10	25	2	100
Professions intermédiaires	27	12	11	48	3	100
Employés	15	10	9	60	7	100
Ouvriers	8	10	11	62	9	100
Étudiants	12	6	34	45	3	100
Retraités	25	20	1	23	32	100
Autres inactifs	15	12	6	49	18	100
Ensemble	20	13	8	44	14	100

$\chi^2 = 1101$
ddl = 32
$p < 0{,}001$
Source : Enquête sur les pratiques culturelles des Français, 1997, DEP/Ministère de la Culture.

Un autre exemple d'ACM

Philippe COULANGEON, « les métamorphoses de la légitimité », ARSS, 2010

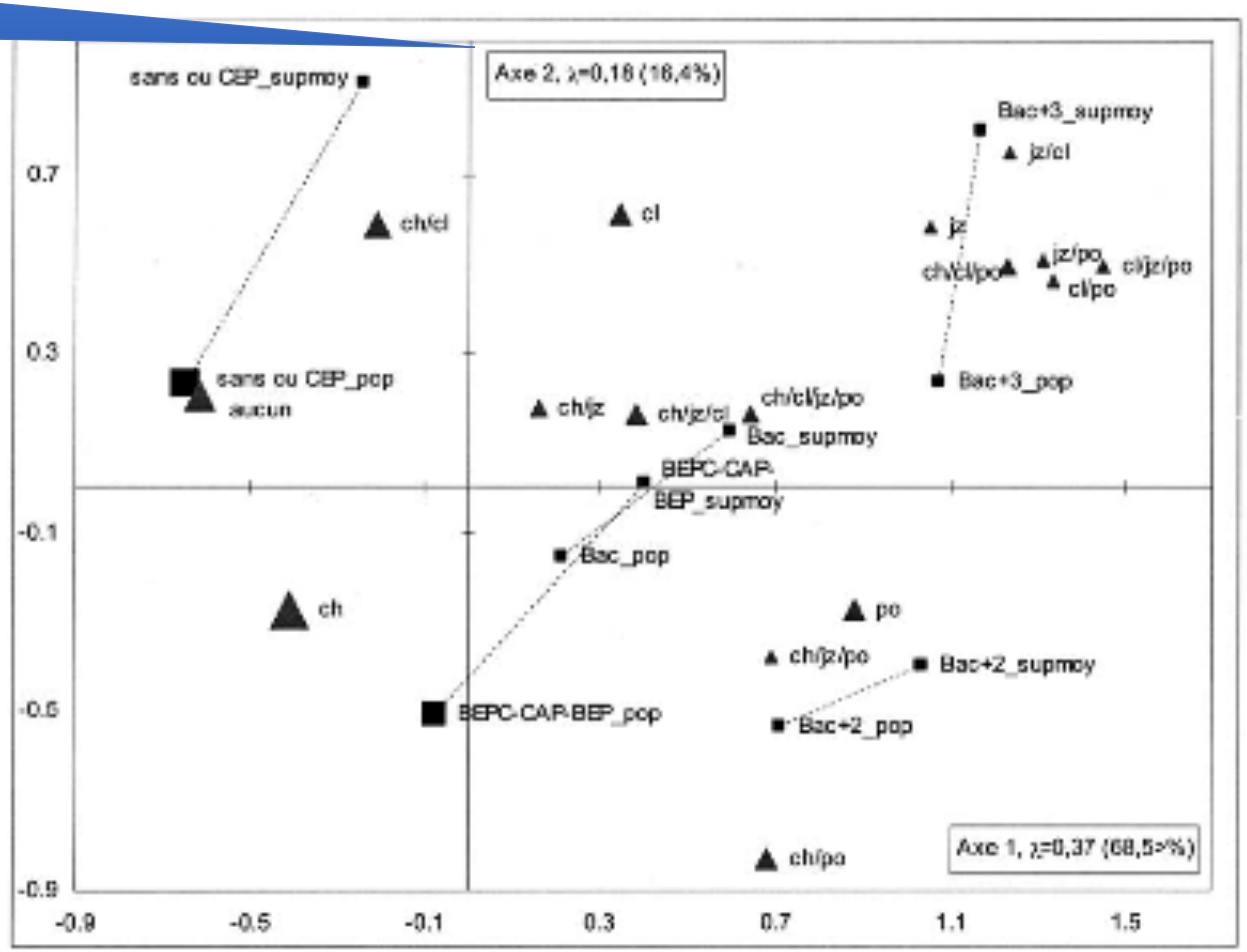
Il part d'un **tableau de contingence** qui croise

- Les combinaisons possibles de genre musicaux musicaux (il y a 4 genres, donc 16 combinaisons possibles)
- Une variable qui combine le niveau de diplôme et l'origine sociale, pour laquelle il retient 10 modalités possibles

La représentation graphique des résultats de cette ACM

Inertie du facteur 2 18.4%

Plan des deux premiers facteurs de l'analyse des correspondances du tableau de contingence des combinaisons de goûts musicaux et des niveaux de diplôme selon l'origine sociale en 2008



Inertie du facteur 1 68.5%

Commentaires du graphique

- Le premier axe oppose
- à un pôle les non diplômés et titulaires du seul certificat d'études d'origine populaire à gauche
- à un autre pôle : les diplômés bac+3 d'origine moyenne ou supérieure. On y trouve aussi les bacheliers, les bac+2 et bac +3 d'origine populaire, mais ils contribuent plus faiblement à la construction du facteur (taille du carré plus petite)
- Conclusion : ce facteur est davantage lié au diplôme qu'à l'origine sociale
- Comment faire le lien avec les goûts musicaux?

Commentaires de la représentation graphique

- Interprétation de l’axe 2

  Pour un même niveau de diplôme on observe que systématiquement quand on se déplace vers le haut, on passe d’une origine populaire à une origine moyenne ou supérieure
- Conclusion : ici l’axe est structuré par l’opposition des individus selon leur origine sociale.

- Quelle correspondance avec les goûts musicaux?

Le lien avec les goûts musicaux

- Sur le premier axe, on repère une opposition entre d'une part le genre « variétés-chansons » et « aucun genre cité »/rock-pop-musique classique : éclectisme
- Sur le deuxième axe, on repère une opposition principale entre « chansons+pop-rock » et « musique classique » ou « musique classique+ chansons » : genres populaires/genres savants

Conclusions de cette ACM

- Constat « proximité de l'origine moyenne et supérieure lorsqu'elle est associée à un niveau de diplôme élevé avec les genres ou combinaisons de genre les plus légitimes. »
- Lecture sociologique :

- « L'impact du niveau de diplôme sur la proximité avec les formes canoniques de la culture légitime paraît ainsi étroitement conditionné par l'origine sociale. »
- « la figure caractéristique de l'omnivore est plus nettement associée à la promotion scolaire des enfants des classes populaires, tandis que les « héritiers » demeurent plus proches de la définition traditionnelle du goût « snob », pour reprendre la terminologie de Peterson, caractérisée par un rapport plus exclusif aux arts savants »
- « faible efficacité de l'institution scolaire, lorsqu'elle n'est pas relayée par la socialisation familiale, dans l'inculcation des normes du goût musical légitime »

Les métamorphoses de la distinction

Taux d'incidence des différents genres cités au titre des genres écoutés le plus souvent selon les groupes sociaux de 1973 à 2008

	1973	1981	1988	1997	2008
Musique classique					
Agriculteurs	8 %	7 %	12 %	10 %	19 %
Patrons de l'industrie et du commerce	18 %	18 %	25 %	14 %	20 %
Cadres sup. et professions libérales	51 %	47 %	57 %	52 %	40 %
Cadres moyens	29 %	26 %	36 %	27 %	26 %
Employés	22 %	16 %	26 %	12 %	18 %
Ouvriers	14 %	7 %	16 %	9 %	16 %
Ensemble	19 %	15 %	26 %	17 %	22 %

Coulangeon, ARSS, 2010

Les métamorphoses de la distinction

Distribution des auditeurs des différents genres au titre des genres écoutés le plus souvent selon les groupes sociaux de 1973 à 2008

	1973	1981	1988	1997	2008
Musique classique					
Agriculteurs	8 %	5 %	3 %	2 %	2 %
Patrons de l'industrie et du commerce	11 %	10 %	7 %	6 %	5 %
Cadres sup. et professions libérales	17 %	23 %	23 %	30 %	27 %
Cadres moyens	16 %	25 %	24 %	22 %	22 %
Employés	19 %	19 %	18 %	17 %	16 %
Ouvriers	29 %	18 %	24 %	23 %	28 %
Ensemble	100 %	100 %	100 %	100 %	100 %

Les métamorphoses de la distinction

- la part des auditeurs de musique classique décline au sein des classes supérieures comme dans d'autres catégories
- Mais ces catégories supérieures restent surreprésentées parmi les mélomanes
- Or le poids de ces catégories dans la population étudiée augmente entre 1973 et 2008
- C'est donc par un effet de structure que la part des mélomanes dans l'ensemble de la population n'a pas baissé et a même légèrement augmenté entre 1973 et 2008 (pas de façon linéaire)

Pourquoi la part des mélomanes a baissé au sein des catégories supérieures?

« les transformations intervenues dans le recrutement social des catégories supérieures au cours de la période couverte par la comparaison des cinq enquêtes, qui affaiblit vraisemblablement l'impact des formes familiales de transmission des habitudes culturelles caractéristiques des héritiers de la culture légitime »

Qui a déclaré dans un entretien au Monde, en décembre 2009 ….?

Emotion musicale

J'ai eu ma période chanson française – Brel, Brassens, Moustaki –, mais ensuite, c'est surtout vers la *pop music* anglaise et américaine que je me suis tourné. J'étais plus Beatles que Rolling Stones (même si j'étais au dernier concert des Stones au Stade de France) et mon premier 33-tours, c'était *Revolver*. Je revois encore la pochette, les visages dessinés au crayon. Et je n'ai plus décroché : Dire Straits, Jefferson Airplane restent parmi mes favoris. En 2006, j'ai surpris une journaliste de télévision en lui faisant découvrir Amy Winehouse.

Dans ma vie d'élu, je suis très fier d'avoir développé le Festival de musique baroque de Sablé-sur-Sarthe, qu'avait créé mon prédécesseur à la mairie, Joël Le Theule. Une de mes plus grandes émotions musicales, je l'ai d'ailleurs vécue à Sablé, au centre culturel : c'était un des derniers concerts du pianiste hongrois Georges Cziffra, un homme hors du commun, un virtuose, mais surtout un artiste d'une extraordinaire sensibilité, et d'une grande générosité. » ■

Propos recueillis par Nathaniel Herzberg