Comment traduire sous forme de probabilités les résultats d'une modélisation logit ?

Jérôme Deauvieau

Bulletin de Méthodologie Sociologique 2010 105: 5
DOI: 10.1177/0759106309352586

Comment traduire sous forme de probabilités les résultats d'une modélisation logit ?

Bulletin de Méthodologie Sociologique
105 5–23

DOI: 10.1177/0759106309352586
http://bms.sagepub.com

Jérôme Deauvieau
Laboratoire Printemps, CNRS/UVSQ

Abstract

How to Translate a Logit Model into Probabilities: It is common in sociology to try to translate the results of a logit model into probabilities or percentages; in other words, into the language of cross tabulations. The purpose of this paper is to present and discuss three ways of performing this operation. This is done using numerical data taken from the FQP 2003 survey by INSEE. We present and discuss successively the method "of deviation from a reference situation", the "experimental deviation" method and finally the method of "pure deviation".

Résumé

Il est courant en sociologie de chercher à traduire les résultats d'une modélisation logit sous la forme de probabilités ou de pourcentages, autrement dit dans le langage du tableau croisé. L'objet de cet article est de présenter et discuter trois façons différentes de réaliser cette opération. On utilise pour cela un exemple numérique concret tiré de l'enquête FQP 2003 de l'INSEE avec lequel sont présentés et discutés successivement la méthode ≪ de l'écart à la situation de référence ≫, la méthode de ≪ l'écart expérimental ≫, et enfin la méthode de ≪ l'écart pur ≫.

Mots clés

Modélisation logit, Régression logistique, Probabilités

Keywords

Logit Models, Logistic Regression, Probabilities

Corresponding author:
Jérôme Deauvieau, Laboratoire Printemps, CNRS / UVSQ
Email: Jerome.Deauvieau@uvsq.fr

Introduction

La régression logistique fait partie des dernières méthodes statistiques importées en sociologie en France. Son introduction dans cette discipline a occasionné deux registres de débats. Le premier a porté sur la légitimité même du raisonnement « toutes choses égales par ailleurs » inhérent aux méthodes de régression multiple, débat qui est d'ailleurs antérieur en sociologie quantitative à la méthode elle-même[1]. Le second est plutôt d'ordre méthodologique puisqu'il porte sur les façons de modéliser lorsqu'on est en présence d'une variable à expliquer catégorielle.[2] Même si la discussion continue de porter sur le principe même de la régression logistique, il semblerait que cette méthode fasse aujourd'hui partie de la boîte à outils du sociologue quantitativiste. L'objet de cet article est de ce fait résolument méthodologique, et vise à présenter et à discuter différentes façons de traduire les résultats d'une modélisation logit sous la forme de probabilités.

En sociologie, on cherche en effet souvent à transformer le résultat d'une modélisation logit en probabilités ou en pourcentages pour au moins deux raisons. La première raison est liée au mode d'utilisation de la régression logistique par les sociologues. Ces derniers utilisent très souvent des variables explicatives catégorielles (sexe, PCS . . .), et lorsqu'ils sont en présence de variables numériques (salaire, âge . . .), l'usage courant veut qu'elles soient mises en catégories. Ce choix en est bien un, puisqu'il est tout à fait possible de laisser les variables numériques dans un modèle de régression logistique. Cette pratique de mise en catégories a bien entendu à voir avec le fait que l'intérêt du sociologue pour les comportements l'amène à croiser plus souvent des variables catégorielles. La statistique du sociologue relève de la catégorie, là où celle de l'économiste relève du nombre.[3] Si la modélisation logit est utilisée seulement avec des variables catégorielles, alors le lien avec l'univers du tableau croisé est évident. Les données de base de la modélisation peuvent en effet être présentées comme un tableau croisé d'une profondeur égale au nombre de variables mises dans le modèle[4]. Dans ces conditions, il est naturel de chercher à exprimer les résultats d'une modélisation logit sous la forme d'une probabilité ou d'un pourcentage ; c'est-à-dire dans le langage du tableau croisé. La seconde raison qui pousse les sociologues à présenter les résultats sous forme de probabilités est probablement d'ordre pédagogique. Ce type de présentation facilite en effet la communication des résultats de la modélisation à un large public. Cette préoccupation est présente dans les revues scientifiques généralistes, mais aussi dans les publications de résultats de la statistique sociale, susceptibles d'intéresser au-delà du cercle des spécialistes de sociologie quantitative.

Or, la présentation des résultats d'un modèle logit sous forme de probabilités ne va pas de soi. Il existe en effet plusieurs méthodes de présentation des résultats sous forme de probabilités dont les logiques et les résultats sont très différents. Nous en présenterons ici trois : l'écart à la situation de référence, l'écart expérimental, et l'écart pur, en donnant à voir leur logique propre sur un exemple concret et en discutant de leurs avantages et inconvénients respectifs.

Lire les résultats d'une modélisation logit

Pour la démonstration, on utilise des données tirées de l'enquête Formation Qualification Professionnelle de l'INSEE réalisée en 2003. La population étudiée est composée des

professions intermédiaires administratives et commerciales des entreprises en 1998 qui sont toujours en emploi en 2003. On cherche à modéliser la probabilité de connaître une mobilité professionnelle entre 1998 et 2003, et on introduit pour cela dans le modèle les trois variables explicatives suivantes : le sexe (code 0: femme ; code 1: homme), l'âge (1: inférieur ou égal à 35 ans, 2: entre 36 et 45 ans, 3: plus de 45 ans), le diplôme (0 : inférieur au bac, 1 : supérieur ou égal au bac).[5] On cherche à modéliser la probabilité de devenir cadre, plus précisément d'appartenir en 2003 au groupe 3 de la nomenclature des PCS ; c'est-à-dire, les cadres et les professions intellectuelles supérieures. L'échantillon sur lequel est appliqué ce modèle est composé des 1.237 individus de l'enquête FQP 2003, qui étaient ≪ professions intermédiaires du privé ≫ en 1998, et qui soit sont restés ≪ professions intermédiaires du privé ≫ (PI), soit sont devenus cadre en 2003. La variable à expliquer est dichotomique, nous sommes donc en présence d'un modèle logit simple (ou dichotomique).

La modélisation va porter non pas directement sur la probabilité de devenir cadre, mais sur le logit de cette probabilité, d'où le nom donné à ce type de modèle. Le logit correspond à la formule suivante :

$$\ln \frac{\Pr(Y = cadre)}{1 - \Pr(Y = cadre)}$$

Ce logit constituera le membre de gauche de l'équation de régression. On trouvera à droite une équation linéaire formée des variables explicatives. Chaque variable explicative introduite dans le modèle est mise sous forme dichotomique, avec un codage en 0 ou 1 (dit ≪ présence ≫ / ≪ absence ≫). Par exemple, pour le sexe, on a choisi de coder les femmes en 0 et les hommes en 1. 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 age 1 (code 1 si âge inférieur à 35 ans, sinon 0), age 2 (code 1 si âge compris entre 35 et 45 ans, sinon 0), age 3 (code 1 si âge supérieur à 45 ans, sinon 0).

Le modèle s'écrit alors :

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

Avec diplôme = 0 si inférieur au bac, 1 si supérieur
Sexe = 0 si femme, 1 si homme
Age 2 = 1 si age compris entre 36 et 45, sinon 0
Age 3 = 1 si âge strictement supérieur à 45 ans, sinon 0

On remarquera qu'il manque la variable dichotomique age1 dans le modèle. L'explication est simple: pour représenter trois situations différentes, deux variables dichotomiques suffisent. Le tableau suivant présente ce raisonnement (Tableau 1). Pour représenter la variable A, qui a trois modalités, l'information contenue dans les deux premières variables (modalités 1 et 2) suffit, puisqu'il est possible de déterminer la valeur prise par la modalité 3 à partir des valeurs prises par les modalités 1 et 2. La dernière variable est en

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

fait une combinaison linéaire des deux précédentes, qui n'apporte donc pas dans le cas présent d'information supplémentaire. C'est la raison pour laquelle on doit omettre une variable dichotomique dans le modèle. Une variable à n modalités est donc toujours représentée par n-1 variables dichotomiques dans le modèle.

Les coefficients B0 à B4 sont les coefficients estimés par le modèle logit. Ce sont ces coefficients qu'il faudra interpréter pour lire le résultat de la modélisation. Le jeu de coefficients et de variables dichotomiques permet de modéliser le logit de la probabilité de devenir cadre dans toutes les situations distinguées par le modèle. Dans notre exemple, il y a 12 situations possibles, correspondant aux 2 catégories de sexe multipliées par les 2 catégories de diplôme et les 3 catégories d'âge. Les résultats du modèle permettent de calculer le « logit » de toutes ces situations en utilisant la combinaison de variables *ad hoc*. Ainsi, le logit des femmes n'ayant pas le bac et ayant moins de 35 ans correspond à une mise à zéro de l'ensemble des variables du modèle. Le logit de cette situation est donc égal au premier coefficient du modèle (B0), qui est le seul coefficient qui ne dépend pas des variables introduites dans le modèle. On appelle cette catégorie *la situation de référence du modèle*. Si on veut calculer le logit de la catégorie des femmes n'ayant pas le bac et ayant entre 35 et 45 ans, il suffit d'ajouter au coefficient de la situation de référence (B0) le coefficient correspondant à cette catégorie d'âge (B3). On continue ainsi pour l'ensemble des 12 situations du modèle. Les résultats sont reportés dans le Tableau 2 en utilisant les valeurs des coefficients estimés par le modèle.

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

Il n'est cependant pas nécessaire de calculer le logit de l'ensemble des situations pour extraire l'information pertinente d'un modèle logit. Chaque coefficient correspond à l'augmentation ou à la diminution du logit lorsque l'on passe du code 0 au code 1 de la variable adossée à ce coefficient. Par exemple, pour le sexe, le modèle indique que le logit de la probabilité de devenir cadre augmente de 0,59 lorsque l'on passe de la situation femme à la situation homme. Cette augmentation, et c'est là un point crucial de l'affaire, est la même « toutes choses égales par ailleurs » ; c'est-à-dire, quelle que soit la configuration des autres variables introduites dans le modèle. En d'autres termes, si on compare le logit des hommes de moins de 35 ans, diplômés du supérieur, au logit des femmes de moins de 35 ans, diplômées du supérieur, on trouvera un écart de 0,59, et il en sera de même si on compare hommes et femmes chez les plus de 45 ans n'ayant pas le bac (on vérifiera aisément cette caractéristique dans Tableau 2).

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.

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.

On peut donc tirer du modèle les informations suivantes :

Le logit des hommes est supérieur à celui des femmes, le logit des diplomés est supérieur à celui des non diplômés. Pour la variable âge, on peut ordonner les modalités comme suit : logit age 3 < logit age 2 < logit age 1. On présente les résultats sous la forme d'un tableau en mettant en face le coefficient associé à la modalité, ce qui permet d'un coup d'œil d'avoir l'ensemble des résultats du modèle (Tableau 3).

Il ne nous reste plus qu'à examiner les liens entre ce « logit » et la probabilité qu'on cherche à expliquer. Partons des deux logit suivants :

$$\text{Logit } 1 = \ln\frac{P1}{1-P1} \text{et Logit } 2 = \ln\frac{P2}{1-P2}$$

Si on pose que le logit 1 est inférieur au logit 2, que peut-on en déduire sur la relation entre P1 et P2 ? Comme la fonction ln est croissante, on déduit de la première relation (logit 1 < logit 2) que :

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

Donc, un logit inférieur à un deuxième logit implique que la probabilité correspondant au premier logit est elle aussi inférieure à la probabilité correspondant au second logit. Le logit des hommes est supérieur au logit des femmes, on en déduit donc directement que la probabilité d'être cadre en 2003 est plus élevée pour les hommes que pour les femmes, ≪ toutes choses égales par ailleurs ≫ ; c'est-à-dire, dans l'ensemble des situations de contraste repérées par le modèle. Il ne reste plus qu'à lire l'ensemble des résultats du modèle en regardant le signe du coefficient et en ordonnant, le cas échéant, l'ensemble des coefficients correspondant à une variable donnée (par exemple, les coefficients liés à la variable âge indiquent que la probabilité de devenir cadre diminue selon l'âge).[6]

La lecture d'un modèle logit à partir des coefficients est somme toute simple, puisqu'il s'agit en fait d'un modèle linéaire avec une variable à expliquer qui a une forme particulière (le logit). En revanche, lorsqu'il s'agit de traduire les résultats d'un modèle logit sous forme de probabilités, les choses se compliquent. Trois façons différentes de présentation en probabilité vont être successivement abordées.

L'écart à la situation de référence

Cette première façon de présenter les résultats d'une modélisation logit sous forme de probabilité est probablement aujourd'hui encore la plus courante en France.[7] Examinons son principe. Nous avons vu qu'il est possible de calculer le logit estimé par le modèle pour l'ensemble des situations repérées par ce modèle. Il est également possible de transformer ce logit en probabilités, à l'aide de la formule de transition suivante :

$$\text{Si } \ln\frac{\Pr(Y=1)}{1-\Pr(Y=1)} = \text{L, alors Pr (Y = 1)} = \frac{1}{1+\exp^{-L}}$$

Avec cette formule nous calculons pour chaque situation (et donc en fait pour chaque individu, étant entendu que chaque individu dans une situation a la même probabilité de devenir cadre) la probabilité estimée par le modèle (Tableau 4).

On pourrait s'en tenir là, mais ce mode de présentation est d'autant plus compliqué à lire qu'il y a de catégories introduites dans le modèle. On simplifie la présentation en sélectionnant une situation parmi les douze possibles. Les différences de probabilité sont calculées pour chaque changement de modalités par rapport à cette situation donnée. La situation de référence est généralement choisie comme situation de départ, étant entendu que ce choix est dicté par le seul impératif de la simplicité du calcul, n'importe quelle autre situation pouvant être sélectionnée. On calcule pour cette situation la probabilité de connaître l'évènement modélisé, ici une probabilité de devenir cadre égale à 0,12. On change ensuite une modalité, par exemple homme à la place de femme pour commencer, et on regarde la probabilité de devenir cadre (ici 0,20). La différence entre les deux probabilités indique l'écart de probabilité entre hommes et femmes dans cette

Tableau 4. Calcul des probabilités prédites par le modèle pour les 12 situations

Situations	Logit	Probabilités estimées par le modèle	Probabilités calculées directement à partir des données
Pas le bac, femme, age 1	−1,95	0,12	0,14
Pas le bac, femme, age 2	−2,24	0,1	0,11
Pas le bac, femme, age 3	−2,58	0,07	0,06
Pas le bac, homme, age 1	−1,36	0,2	0,18
Pas le bac, homme, age 2	−1,64	0,16	0,17
Pas le bac, homme, age 3	−1,99	0,12	0,12
Bac, femme, age 1	−1,20	0,23	0,24
Bac, femme, age 2	−1,48	0,18	0,14
Bac, femme, age 3	−1,83	0,14	0,16
Bac, homme, age 1	−0,60	0,35	0,35
Bac, homme, age 2	−0,89	0,29	0,28
Bac, homme, age 3	−1,23	0,23	0,32

Source: INSEE, enquête FQP, 2003.

Tableau 5. Expliquer le passage à la catégorie cadre

	Probabilité passage cadre
Probabilité à la situation de référence	0,12
Sexe	
Femme	*0,12*
Homme	0,20
Diplôme	
Inférieur au bac	*0,12*
Supérieur au bac	0,33
Age	
Age 1	*0,12*
Age 2	0,10
Age 3	0,07

Source: INSEE, enquête FQP, 2003.

situation. Il ne reste plus qu'à procéder de même pour toutes les modalités du modèle, et à inscrire les résultats dans un tableau (Tableau 5).

L'intérêt de ce mode de présentation est double. D'une part, il est relativement simple à comprendre ; d'autre part, il a l'avantage d'être très proche d'une logique de tableau croisé, puisqu'on peut même comparer directement les résultats en probabilité de la modélisation avec les résultats issus des données réelles calculés à l'aide de tableaux croisés (dernière colonne du Tableau 4). Mais sa simplicité apparente peut se révéler problématique. En effet, il est tentant de regarder la taille des effets, et de les commenter. Or, le risque évident consiste à généraliser une différence de probabilités entre deux modalités à l'ensemble des situations. Prenons, par exemple, la différence entre hommes et femmes. Nous avons vu que l'écart de logit est positif, donc l'écart en probabilité sera

Tableau 6. Différence de logit et différences de probabilités entre hommes et femmes

Situation	résultat logit homme	résultat logit femme	probabilité hommes	probabilité femmes	écart logit	différence des probabilités
Pas le bac, age1	−1,36	−1,95	0,20	0,12	0,59	0,08
Pas le bas, age 2	−1,65	−2,25	0,16	0,09	0,59	0,07
Pas le bac, age 3	−1,99	−2,58	0,12	0,07	0,59	0,05
Bac, age 1	−0,61	−1,20	0,35	0,23	0,59	0,12
Bac, age 2	−0,90	−1,49	0,29	0,18	0,59	0,11
Bac, age 3	−1,24	−1,83	0,22	0,14	0,59	0,09

Source: INSEE, enquête FQP, 2003.

nécessairement positif entre les hommes et les femmes dans toutes les situations (en l'occurrence, dans les six situations possibles). Mais la présentation en écart à la situation de référence donne une grandeur pour cet écart ; et la tentation est grande de la considérer comme absolue. Cette erreur consiste alors à énoncer la proposition suivante : « l'écart entre hommes et femmes est de 8 points de pourcentages, toutes choses égales par ailleurs », ce qui est faux. On le vérifie aisément en calculant l'écart de probabilités entre hommes et femmes dans toutes les situations (Tableau 6).

La différence des probabilités de devenir cadre entre hommes et femmes varie de 0,05 à 0,12, selon les situations. On touche là à la limite de ce mode de présentation des résultats. En effet, cette transformation du résultat du modèle en probabilités n'a pas du tout la même propriété que la différence de logit qui reste stable selon la situation et dont on peut commenter le résultat de façon absolue. C'est la raison pour laquelle cette façon de convertir l'écart de logit en probabilité ne peut être assimilée à une *traduction* des résultats du modèle, il s'agit bien plus d'une *illustration* des résultats du modèle à un endroit particulier de l'échantillon.

L'écart expérimental[8]

Il existe une deuxième façon de transformer les résultats du modèle sous forme de probabilité, qu'on appellera « l'écart expérimental ». Le principe est le suivant. On calcule pour chaque individu de l'échantillon sa probabilité individuelle de connaître l'évènement modélisé.[9] A partir de ces probabilités individuelles, on calcule des probabilités « théoriques » en assumant la posture « expérimentale » inhérente aux méthodes de régression multiple. En effet, la base épistémologique de la régression multiple consiste à découper les individus selon leurs différentes caractéristiques sociales, et à mesurer l'effet d'une des caractéristiques indépendamment des autres. Cela revient en fait à imiter le raisonnement expérimental qu'on trouve, par exemple, en biologie : je cherche à mesurer l'effet d'une caractéristique sur un phénomène donné, et pour cela je fais une expérience consistant à donner cette caractéristique à un groupe, à omettre cette caractéristique pour un second groupe, et à regarder ce qui se passe en comparant les deux groupes. Transposé en sciences sociales, ce raisonnement consisterait, par exemple, à attribuer un sexe féminin à un premier groupe, à attribuer un sexe masculin à un second groupe, et à vérifier si le premier groupe a un salaire inférieur au second.[10] C'est le même

principe qui guide les méthodes de régression, et qui constitue d'ailleurs la principale critique adressée à ces méthodes : en cherchant à raisonner « toutes choses égales par ailleurs », la régression produit un fort effet déréalisant puisque chacun sait que dans la réalité sociale les choses sont souvent précisément « inégales par ailleurs ».[11]

Le jeu de coefficients estimés par le modèle permet précisément de se mettre, par le calcul, dans une posture expérimentale. Prenons, par exemple, la variable sexe dans notre exemple. Le modèle indique que les femmes ont un logit inférieur de 0,59 à celui des hommes. Pour traduire cet écart en probabilités, il suffit de réaliser l'expérience suivante sur l'échantillon : si tous les individus de mon échantillon étaient des femmes, quelle serait la probabilité moyenne de devenir cadre au bout de cinq ans ? Deuxième manipulation, si tous les individus de mon échantillon étaient des hommes, quelle serait la probabilité moyenne de devenir cadre au bout de cinq ? Il suffit ensuite de faire la différence entre ces deux probabilités, et on obtient ainsi une mesure en probabilité de la différence entre hommes et femmes de devenir cadre au bout de cinq ans, « toutes choses égales par ailleurs ».

Concrètement, il suffit de calculer pour chaque individu de l'échantillon, la probabilité de devenir cadre en appliquant les deux équations suivante :

Premier cas : l'échantillon est de façon expérimentale constitué exclusivement de femmes, on omet donc pour tous les individus le coefficient lié au sexe. On calcule la probabilité de chaque individu de devenir cadre à partir de la formule suivante :

$$\mathrm{P}(\mathrm{Y} = \text{cadre}) = \frac{1}{1 + \exp^{-(-1{,}96 + 0{,}75 * diplome - 0{,}29 * age\ 2 - 0{,}63 * age\ 3)}}$$

Deuxième cas : l'échantillon est constitué cette fois exclusivement d'hommes, on ajoute donc pour chaque individu de l'échantillon (quel que soit son sexe) le coefficient lié au sexe dans l'équation. La formule de calcul de la probabilité individuelle devient alors :

$$\mathrm{P}(\mathrm{Y} = \text{cadre}) = \frac{1}{1 + \exp^{-(-1{,}96 + 0{,}75 * diplome + 0{,}59 - 0{,}29 * age\ 2 - 0{,}63 * age\ 3)}}$$

En appliquant dans notre exemple les deux formules successivement sur l'ensemble de l'échantillon, on trouve dans le premier cas une probabilité de 0,12, dans le second cas de 0,20, la différence de 0,08 point donne l'écart en probabilité de devenir cadre entre hommes et femmes, « toutes choses égales par ailleurs » ; c'est-à-dire, si je ne change que cette caractéristique dans l'échantillon. Autrement dit, si tous les individus étaient des hommes et qu'ils gardaient les mêmes caractéristiques par ailleurs, leur probabilité moyenne de devenir cadre serait de 0,20. Si tous les individus étaient des femmes et gardaient les mêmes caractéristiques par ailleurs, leur probabilité moyenne de devenir cadre serait de 0,12. L'écart entre ces deux probabilités mesure l'effet net du sexe sur la probabilité de devenir cadre. On refait la même manipulation pour l'ensemble des variables et on indique à chaque fois le résultat dans un tableau (Tableau 7). L'avantage évident de la présentation des probabilités sous la forme de l'écart expérimental est qu'elle permet de s'affranchir d'une situation donnée, puisqu'on observe ainsi un écart en probabilités qui vaut en « moyenne ». Il s'agit donc véritablement d'une méthode de *traduction* des

Tableau 7. Présentation de résultats avec l'écart expérimental

	probabilité de devenir cadre
Probabilité moyenne	0,16
Sexe	
Femme	0,12
Homme	0,2
Diplôme	
Inférieur au bac	0,13
Supérieur au bac	0,23
Age	
Age 1	0,2
Age 2	0,16
Age 3	0,12

Source: INSEE, enquête FQP, 2003.

résultats de la modélisation et non seulement d'*illustration* de ces mêmes résultats comme l'écart à la situation de référence.

L'inconvénient principal de ce mode de présentation est que le résultat obtenu n'est pas forcément égal au contraste logistique tel qu'il est indiqué par le coefficient logit du modèle. Ainsi par exemple, le coefficient logit associé au diplôme vaut 0,75, d'après le modèle. En appliquant les probabilités trouvées avec l'écart expérimental (0,23 pour les diplômés et 0,13 pour les non diplômés), on obtiendrait alors un coefficient associé de 0,70 (en faisant le calcul suivant :

$$\ln\frac{0,23}{0,77} - \ln\frac{0,13}{0,87}).$$

L'écart « pur »

La troisième façon de présenter les résultats sous forme de probabilités (l'écart « pur ») respecte quant à elle la valeur des coefficients estimés par le modèle.[12] Le principe consiste à trouver les probabilités associées aux modalités d'une variable explicative qui satisfassent aux deux conditions suivantes :

1. l'écart entre les probabilités respecte le contraste logistique entre les modalités tel qu'il est défini par le coefficient du modèle ;
2. la moyenne pondérée des probabilités liées aux modalités de la variable explicative est égale à la probabilité moyenne de la variable à expliquer sur l'ensemble de l'échantillon.

La première contrainte revient tout simplement à traduire directement le coefficient logit sous la forme de probabilités. Il s'agit donc de trouver les probabilités P1 et P2 tel que :

$$\frac{\dfrac{P1}{1-P1}}{\dfrac{P2}{1-P2}} = \exp \text{ (coefficient du modéle donnant le contraste entre P1 et P2).}$$

Cette équation admet une infinité de solutions. Ainsi par exemple, si le coefficient logit vaut 0,75, alors on peut trouver plusieurs jeux de probabilités P1 et P2, comme les couples (0,7 pour P1 et 0,52 pour P2), ou encore (0,6 ; 0,41). L'ajout d'une deuxième contrainte permet de poser un système d'équation qui n'admet plus qu'une et une seule solution[13]. Cette seconde contrainte consiste à dire que la moyenne pondérée des probabilités P1 et P2 doit être égale à la probabilité moyenne sur l'ensemble de l'échantillon.[14]

Prenons en guise d'exemple le cas de la variable sexe dans le modèle logit dichotomique que nous avons utilisé jusqu'à présent. Nous savons que le contraste logistique entre hommes et femmes vaut, selon le modèle, 0,59, donc que :

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

avec P1 probabilité de devenir cadre pour les hommes et P2 probabilité de devenir cadre pour les femmes.

On transforme cette égalité sous la forme d'un rapport des chances en utilisant les propriétés de la fonction ln:

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

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

On tire donc des résultats de la modélisation que le rapport des chances entre hommes et femmes vaut 1,80. Ce sera donc la première contrainte qu'il faudra respecter : trouver P1 (hommes) et P2 (femmes) tel que le rapport des chances entre hommes et femmes sera égal à 1,8.

La deuxième contrainte consiste à trouver P1 et P2 tel que la moyenne pondérée de ces deux probabilités soit égale à la probabilité de devenir cadre dans l'ensemble de l'échantillon. Comme on a dans nos données 610 hommes et 627 femmes, et que la probabilité de devenir cadre est de 0,16 sur l'ensemble de l'échantillon, on obtient comme seconde contrainte l'équation suivante :

$$\frac{610P1 + 627P2}{1237} = 0,16$$

Nous sommes en présence d'un système de deux équations à deux inconnues qu'il est possible de résoudre pour trouver les valeurs de P1 et P2.[15] Les résultats pour l'ensemble des modalités sont présentés dans le Tableau 8.

Comment lire ces résultats ? Nous dirons que la traduction du coefficient logit lié à la variable sexe produit une probabilité pour les hommes de 0,2 et une probabilité pour les femmes de 0,12, soit un écart pur de 0,08 point. Ce mode de présentation est donc une traduction directe du coefficient, et tout comme l'écart expérimental (et contrairement à l'écart à la situation de référence), il ne dépend pas d'une situation donnée mais est

Tableau 8. Présentation des résultats avec l'écart pur

	Probabilité de devenir cadre
Probabilité moyenne	0,16
Sexe	
Femme	0,12
Homme	0,20
Diplôme	
Inférieur au bac	0,13
Supérieur au bac	0,24
Age	
Age 1	0,21
Age 2	0,16
Age 3	0,12

Source: INSEE, enquête FQP, 2003.

valable en tout point de l'échantillon. Il s'agit donc bien également d'une façon de *traduire* les résultats d'une modélisation logit sous la forme de probabilités.

Conclusion

Quel mode de présentation en probabilité vaut-il mieux adopter ? Tout dépend de l'objectif de la présentation des résultats. S'il s'agit de faire une présentation simplifiée – par exemple, pour un public non averti – la meilleure des solutions consiste probablement à s'en tenir au signe des effets. On indique ainsi en face de chaque modalité le signe du coefficient logit, et on peut ainsi dire que « toutes choses égales par ailleurs », la modalité A fait augmenter ou baisser la probabilité de connaître l'évènement par rapport à la modalité B. Si on souhaite vraiment traduire en probabilité les résultats du modèle, la présentation sous forme d'écart à la situation de référence peut sembler très pédagogique, mais elle porte un sérieux risque de confusion en indiquant un écart en probabilité qui est propre à une situation donnée. Le risque est grand pour le lecteur de considérer l'écart en probabilités comme valable en tout point de l'échantillon. L'erreur devient encore plus grave si le lecteur compare la taille des effets de deux variables et indique que telle variable a un effet plus fort puisque l'écart observé est plus élevé que pour telle autre variable, alors même que cette hiérarchie peut tout à fait être inversée à un autre point de l'échantillon.

En conséquence, pour traduire les résultats d'un modèle logit sous forme de probabilités, il vaut mieux utiliser l'écart expérimental ou l'écart pur. Les résultats obtenus peuvent bien entendu différer entre ces deux façons de faire puisque les principes de calcul sont très différents. Dans le cas de l'écart expérimental, l'opération consiste à prendre le coefficient logit calculé par le modèle et à le plonger dans l'échantillon pour faire une expérience « toutes choses égales par ailleurs ». Dans le cas de l'écart pur, il s'agit de prendre le coefficient logit et de le traduire sous forme de probabilité en respectant la valeur du coefficient logit et la probabilité brute moyenne de la variable expliquée. L'échantillon dans ce cas intervient pour calculer le coefficient logit, mais plus ensuite pour traduire ce coefficient sous forme de probabilités. Ces deux méthodes ont leur

logique propre et permettent donc d'examiner la traduction du coefficient logit sous des angles différents. Parce que les résultats obtenus peuvent différer, l'utilisation simultanée de ces deux méthodes permet de souligner qu'il n'existe pas une façon naturelle et unique de transformer les résultats d'un modèle logit en probabilités.

ANNEXE 1

On donne en annexe l'ensemble du programme permettant sous SAS de retrouver les calculs effectués dans cet article (à l'exception de l'écart pur qui n'est pas programmé sous SAS). On peut trouver auprès du Centre Maurice Halbwachs (CMH) la base de données de l'enquête FQP 2003 de l'INSEE qui a servi pour les exemples. Le logiciel gratuit Trideux calcule directement les trois façons de présenter les résultats d'une modélisation logit sous la forme de probabilités.

```
libname FQP03 "G:\bases\FQP\fqp03new";

/*construction des tables de départ*/

DATA PIprivA;
set fqp03.Qi03;
if occu98=1 and occu03=1;
if csa=46;
run;

/*************************************************************************************
*********************************/
/*population: les PI de 1998 devenus cadres, PI,employés ou ouvriers en 2003. Recodage des
variables explicatives*/

data PIprivB;
set PIprivA;

if 3<=GStot<=6;

if GStot=3 then arrive=1;
if GStot=4 then arrive=3;
if 5<=GStot<=6 then arrive=2;

agefqp=ag*1;

if 20<=ageFQP<=35 then ageA=1;
if 35<ageFQP<=45 then ageA=2;
if ageFQP > 45 then ageA=3;

if 4<=ddipl<=7 then diplome=0;
if 1<=ddipl<=3 then diplome=1;
if ddipl=. then diplome=0;

if S=1 then sexe=1;
if S=2 then sexe=2;
```

```
if ageA=1 then ageA1=1; else ageA1=0;
if ageA=2 then ageA2=1; else ageA2=0;
if ageA=3 then ageA3=1; else ageA3=0;

if sexe=1 then sexe1=1; else sexe1=0;
run;

/*********************************************************************/
/* on ne garde que ceux qui en 2003 sont cadres ou PI*/
data PIprivC;
set PIprivB;

if arrive=1 or arrive=3;
if arrive=1 then cadre=1; else cadre=2;
run;

proc freq data=PIprivB;
table arrive*GSTOT ageA*ageA1*ageA2*ageA3 diplome*ddipl sexe*sexe1/list missing;
run;

/* modèle logit dichotomique*/
proc logistic data=PIprivC outest=sor0;
model cadre = diplome sexe1 /*ageA1*/ ageA2 ageA3;
run;

/*********************************************************************/
/* calcul des sorties en probabilité avec écart à la situation de ref*/

data proba0; set sor0;
qinterc=1/(1+exp (-intercept));
zdiplome=-(intercept+diplome); qdiplome=1/(1+exp(zdiplome));
zsexe1=-(intercept+sexe1); qsexe1=1/(1+exp(zsexe1));
zageA2=-(intercept+ageA2); qageA2=1/(1+exp(zageA2));
zageA3=-(intercept+ageA3); qageA3=1/(1+exp(zageA3));
run;

proc print data=proba0;
var qinterc qdiplome qsexe1 qageA2 qageA3;
run;

/*********************************************************************/
/*Calcul des sorties en probabilité selon la méthode de l'écart expérimental*/

data probaexp0;
set PIprivC;

expP0= exp ( - (-1.95 + 0.75*diplome + 0.59*sexe1 - 0.29*AgeA2 - 0.63*ageA3));
P0= 1 / (1 + expP0);

/* l'effet du sexe*/
```

```
expfemme= exp ( - (-1.95 + 0.75*diplome - 0.29*AgeA2       - 0.63*ageA3));
exphomme= exp ( - (-1.95 + 0.75*diplome + 0.59 - 0.29*AgeA2 - 0.63*ageA3));

Pfemme= 1/(1 + expfemme);
Phomme= 1/ (1 + exphomme);

/* l'effet du diplôme */

expdip0= exp ( - (-1.95       + 0.59*sexe1 - 0.29*AgeA2 - 0.63*ageA3));
expdip1= exp ( - (-1.95 + 0.75 + 0.59*sexe1 - 0.29*AgeA2 - 0.63*ageA3));

Pdip0= 1/(1 + expdip0);
Pdip1= 1/(1 + expdip1);

/* l'effet de l'âge */

expage1= exp ( - (-1.95 + 0.75*diplome + 0.59*sexe1 ));
expage2= exp ( - (-1.95 + 0.75*diplome + 0.59*sexe1 - 0.28 ));
expage3= exp ( - (-1.95 + 0.75*diplome + 0.59*sexe1 - 0.63 ));

Page1= 1 / (1 + expage1);
Page2= 1 / (1 + expage2);
Page3= 1 / (1 + expage3);

run;

proc means data=probaexp0;
var P0 Pfemme Phomme Pdip0 Pdip1 Page1 Page2 Page3;
run;
```

ANNEXE II[16]

Résolution des systèmes pour le calcul de l'écart pur

Cas général à 2 inconnues. Le système s'écrit alors :

$$\begin{cases} \dfrac{p_2}{1-p_2} : \dfrac{p_1}{1-p_1} = A \\ \dfrac{m_1 p_1 + m_2 p_2}{m_1 + m_2} = B \end{cases}$$

La même démarche permet d'obtenir le système suivant :

$$\begin{cases} p_2 = \dfrac{(m_1 + m_2)B - m_1 p_1}{m_2} \\ (A-1)m_1 p_1^2 + ([(m_1 + m_2)B + m_1 + Am_2 - AB(m_1 + m_2)]p_1 - (m_1 + m_2)B = 0) \end{cases}$$

Le calcul littéral des racines de cette dernière équation est inenvisageable. Seul importe le fait qu'elle ait une solution dans l'intervalle]0 ; 1[, et que la valeur de p_2 associée soit aussi dans]0 ; 1[.

On doit d'abord écarter le cas particulier A = 1 car dans ce cas, l'équation n'est pas du second degré. C'est alors très simple : A = 1 donne immédiatement $p_1 = p_2 = B$.

Pour A ≠1, l'équation est de la forme : $T(p_1) = ap_1^2 + bp_1 + c = 0$.

$T(0) = -(m_1+m_2)B < 0$. $T(1) = (A-1)m_1+(m_1+m_2)B+m_2+Am_1-AB(m_1+m_2)-(m_1+m_2)B = A(1-B)(m_1+m_2) > 0$. (car A > 0 et B < 1 et $m_1+m_2 > 0$).

Cela prouve que le trinôme du second degré T change de signe entre 0 et 1, il a donc deux racines dont l'une dans l'intervalle]0 ; 1[.

En prenant pour p_1 cette racine dans [0 ; 1[, il reste à prouver que la valeur de p_2 associée est dans]0 ; 1[. On reporte la valeur de q dans la première équation :

$$\frac{p_2}{1-p_2} : \frac{p_1}{1-p_1} = A \Leftrightarrow p_2(1-p_1) = Ap_1(1-p_2) \Leftrightarrow p_2(1-p_1+Ap_1)$$
$$= Ap_1 \Leftrightarrow p_2 = \frac{Ap_1}{Ap_1+1-p_1}$$

Or 1-p_1 > 0, donc $Ap_1 + 1 - p_1 > Ap_1 > 0$, cela prouve $0 < p_2 < 1$.

Cas général à n inconnues (n ≥ 2). Le problème est le suivant : n∈N, n≥2. On a (n−1) réels $A_2, A_3, \ldots A_n$ strictement positifs, un réel B dans]0 ; 1[, et n entiers naturels $m_1, m_2, \ldots m_n$. On posera $M = \sum_{i=1}^{n} m_i$. On a n probabilités inconnues (toutes différentes de 0 et 1) qui vérifient le système à n équations :

$$\begin{cases} \omega_i(x) = A_i x/(A-1)x + 1 \; si \; A_i \neq 1 \\ x \; xi \; A_i = 1 \end{cases}$$

On va démonter que ce système admet toujours une et une seule solution.

Considérons la première équation. Elle permet d'exprimer p_2 en fonction de p_1. Le calcul a déjà été fait, on trouve :

$p_2 = \frac{A_2 p_1}{A_2 p_1 + 1 - p_1}$. (Cas particulier : si $A_2 = 1, p_2 = p_1$). Considérons la fonction Φ_2 définie sur [0 ; 1] par : $\varphi_2(x) = \frac{A_2 x}{(A_2-1)x+1}$. Cette fonction rationnelle est bien définie sur [0 ; 1] car $(A_2-1)x+1$ s'annule pour $x = 1/(1-A_2)$ et ce réel n'est pas dans [0 ; 1], car si A_2 est strictement supérieur à 1, il est strictement négatif, et si A_2 est compris strictement entre 0 et 1, il est strictement supérieur à 1. La fonction Φ_2 est donc dérivable sur [0 ; 1] et $\varphi_2'(x) = \frac{A_2}{((A_2-1)x+1)^2} > 0$. La fonction Φ_2 est donc continue strictement croissante sur [0 ; 1] et $\Phi_2(0) = 0$ et $\Phi_2(1) = 1$. Pour le cas particulier $A_2 = 1$, on pose $\Phi_2(x) = x$, et le résultat précédent reste valable.

On étend cette notation aux autres équations du système en posant pour i = 2, 3, ... n:

$p_i = \Phi_i(p_1)$ avec $\varphi_i(x) = \begin{cases} \dfrac{A_i x}{(A_i - 1)x + 1} & si\ A_i \neq 1 \\ x & si\ A_i = 1 \end{cases}$. Le système s'écrit alors :

$$\begin{cases} Pour\ i = 2, 3 \ldots n : p_i = \varphi_i(p_1) \\ \sum_{i=1}^{n} = m_i p_i = M.B \end{cases}$$

On reporte les valeurs de p_2, p_3, ...p_n dans la dernière équation en posant en outre $\Phi_1(x) = x$:

$$\sum_{i=1}^{n} m_i \varphi_i(p_1) - MB = 0. (*)$$

Il est inutile de mettre cette équation sous forme polynômiale comme on a fait pour deux inconnues car le calcul serait long et aboutirait à une équation de degré n.

Considérons la fonction F définie sur [0 ; 1] par : $F(x) = \sum_{i=1}^{n} m_i \varphi_i(x) - MB$

Comme les coefficients m_i sont strictement positifs, les fonctions $m_i\Phi_i$ sont toutes continues strictement croissantes sur [0 ; 1], et donc F est continue strictement croissante sur [0 ; 1]. Comme F(0) = −MB < 0 (car M>0 et B>0) et F(1) = M – MB = M(1 – B) >0 (car M>0 et B<1), la fonction F change de signe sur [0 ; 1], et elle s'annule une et une seule fois dans]0 ; 1[. L'équation (*) a donc une unique solution p_1 dans]0 ; 1[, et les (n-1) premières équations du système donnent l'unique valeur dans]0 ; 1[de chacune des autres inconnues p_2, p_3, ... p_n.
Le système proposé a une et une seule solution.
La résolution pratique du système est très simple : On écrit l'équation (*), on calcule sa solution dans]0 ; 1[par dichotomie par exemple, on calcule ensuite p_2, p_3, ...p_n.

Notes

1. Les débats sur cette question sont à la fois nombreux et anciens, et les références sur le sujet sont trop importantes pour être toutes citées. Pour une critique vigoureuse et antérieure à la régression logistique de l'idée de comparer ≪ toutes choses égales par ailleurs ≫, voir Halbwachs, 1944, et les commentaires qu'en propose Olivier Martin (1999). Sur les débats récents, voir Passeron (1991), Nétumières (1997), Cibois (1999), Desrosières (2001) et Vallet (2006).
2. Ainsi par exemple les discussions dans le *BMS* sur l'opportunité d'utiliser un modèle linéaire ou un modèle logit lorsqu'on modélise une variable qualitative. Voir à ce sujet les articles de Philippe Cibois (1999 et 2000), Aris et Hagenaars (2000), et Aris et Aris (2002).
3. Cf. Desrosières, (2001).
4. On peut aller plus loin en remarquant que dans ce cas la modélisation logit n'est qu'une présentation particulière d'une analyse log linéaire qu'on utilise habituellement pour modéliser les interactions entre des variables catégorielles. En effet, les résultats d'une modélisation logit

peuvent être retrouvés sans difficulté à partir des résultats d'une analyse log linéaire menée sur les mêmes variables. Sur cette question, voir Demaris (1992) et Vallet (2005).

5. On trouvera en annexe les programmes SAS utilisés dans cet article. Le logiciel libre Trideux calcule directement les trois façons de présenter les résultats sous forme de probabilités développées dans cet article.
6. Comme on travaille sur un échantillon, on prendra garde à la valeur de la probabilité associée au test statistique, qui correspond au test de la nullité du coefficient. Plus la probabilité est élevée, plus il y a de chances que le coefficient soit en fait nul dans la population, et donc que l'écart entre les deux catégories associé au coefficient soit nul également. On n'interprète pas dans ce cas le coefficient estimé par le modèle.
7. Elle a notamment été utilisée dans l'article célèbre de Louis André Vallet et Jean-Paul Caille (1995) sur la scolarité des étrangers. La notoriété de cet article a très probablement joué un grand rôle dans l'introduction de la régression logistique en sociologie en France.
8. On trouvera notamment ce mode de présentation des résultats pour le logit multinomial dans Asfa Essafi (2003).
9. Nous avons déjà réalisé ce calcul dans le Tableau 4 à propos de la présentation des résultats sous formes d'écart à la situation de référence. Remarquons qu'en calculant la moyenne des probabilités individuelles prédites par le modèle pour chaque individu, on retrouve la probabilité brute calculée directement sur les données.
10. Sur cette question, voir Nétumières (1997), Behaghel (2006).
11. François Héran note ainsi sur le raisonnement en termes de regression 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).
12. Cette façon de présenter les résultats a été développée par Laurent Toulemon (Léridon et Toulemon, 1997).
13. La démonstration est présentée en Annexe II.
14. Cela revient à fixer le rapport des chances entre P1 et P2 à un niveau qui respecte l'une des données de la réalité, à savoir la probabilité moyenne qu'on cherche à modéliser. Il doit être possible de trouver d'autres types de contraintes en respectant cet esprit.
15. Voir Annexe II. Il est possible de démontrer que ce système avec ses contraintes associées admet toujours une seule solution. La démonstration est présentée en Annexe II, ainsi que la résolution du système pour la variable sexe.
16. L'Annexe II a été rédigée par Roland Deauvieau.

Références

Aris, E et Hagenaars, J (2000) Remarques sur la comparaison entre les modèles linéaire et logit. *Bulletin de Méthodologie Sociologique*, 66: 5–12.

Aris, H et Aris, E (2002) L'analyse des données tabulaires avec les modèles linéaires et log-linéaires. *Bulletin de Méthodologie Sociologique*, 74: 5–32.

Asfa Essafi, C (2003) *Les modèles logit polytomiques non ordonnés : Théorie et applications.* Paris: INSEE, Série des Documents de Travail, Méthodologie Statistique, n. 0301.

Behaghel, L (2006) *Lire l'économétrie*. Paris : La Découverte.

Cibois, P (1999) Modèle linéaire contre modèle logistique en régression sur données qualitatives. *Bulletin de Méthodologie Sociologique*, 64: 5–24.

Cibois, P (2000) Observation et modèle linéaire ou logistique : Réponse à Aris et Hagenaars. *Bulletin de Méthodologie Sociologique*, 67: 54–64.

Demaris, A (1992) *Logit Modeling. Pratical Applications*. Sage University Papers series Quantitative Applications in the Social Sciences, 07–086. Newbury Park, CA: Sage.

Desrosières, A (2001) Entre réalisme métrologique et conventions d'équivalences : Les ambiguïtés de la sociologie quantitative. *Genèse*, 43: 112–127.

Halbwachs, M (1944) *La statistique en sociologie – La statistique. Ses applications. Les problèmes qu'elles soulèvent.* Paris: PUF.

Héran, F (1996) L'école, les jeunes et les parents : Approches à partir de l'enquête éducation [présentation du numéro spécial]. *Economie et statistique*, 296: 5–15.

Léridon, H et Toulemon, L (1997) *Démographie. Approche statistique et dynamique des populations*. Paris: Economica.

Martin, O (1999) Raison statistique et raison sociologique chez Maurice Halbwachs. *Revue d'histoire des sciences humaines*, 1: 69–101.

Nétumières, F (1997) Méthodes de régression et analyse factorielle. *Histoire et Mesure*. 12, 3–4: 271–298.

Passeron, J-C (1991) *Le raisonnement sociologique. L'espace non poppérien du raisonnement naturel*. Paris: Nathan.

Vallet L-A (2005) Utiliser le modèle log linéaire pour mettre à jour la structure du lien entre les deux variables d'un tableau de contingence : Un exemple d'application à la mobilité sociale. *Actes des journées de méthodologie statistique*, Paris: INSEE.

Vallet, L-A (2006) Sur l'analyse de régression en sociologie. Communication au RT20 au congrès de l'Association Française de Sociologie, Bordeaux.

Vallet, L-A et Caille, J-P (1995) Les carrières scolaires au collège des élèves étrangers ou issus de l'immigration. *Éducation et Formations*, 40: 5–14.