Bulletin de Méthodologie Sociologique
2019, Vol. 142 7–31

DOI: 10.1177/0759106319834657
journals.sagepub.com/home/bms

Comparer les résultats d'un modèle logit dichotomique ou polytomique entre plusieurs groupes à partir des probabilités estimées

Jérôme Deauvieau
Centre Maurice Halbwachs (CMH), Département de sciences sociales, École normale supérieure, EHESS, CNRS, PSL University, Paris, France

Abstract

Comparing the results of a multinomial or binary logit models between several groups based on estimated probabilities. In sociology it is common to attempt to compare the effects of a variable within a given logit model conducted with several groups. The most common way of doing this consists in directly comparing the value of the coefficient of the variable in question. However, this practice raises serious methodological problems in the case of a logit model. This point was raised in the sociological literature by Allison in the late 1990s. An increasing number of studies have been dedicated to the subject since the late 2000s. Our objective here is to cover this question for a case in which the variable of interest is categorical and in a general framework combining binary and multinomial logistic mobilisations. The solutions discussed in this article consist in translating a logit coefficient in the form of probabilities. Having presented the issue at stake and the different forms of solution proposed, we study the operationalisation of two methods for the translation of this logit coefficient, experimental deviation and pure deviation, from the point of view of comparison between several groups of results within a logit model.

Corresponding Author:
Jérôme Deauvieau, ENS Campus Jourdan, Centre Maurice Halbwachs (CMH), Département de sciences sociales, École normale supérieure, EHESS, CNRS, PSL University, 75014, Paris, France
Email: Jerome.deauvieau@ens.fr

Résumé
Il est très fréquent en sociologie de chercher à comparer les effets d'une variable d'un même modèle logit réalisé sur plusieurs groupes. La manière la plus courante de réaliser cette opération a souvent consisté à comparer directement la valeur du coefficient logit de la variable d'intérêt. Or, cette pratique soulève de sérieuses objections méthodologiques dans le cas de la modélisation logit. Ce point a été mis en évidence dans la littérature sociologique par Allison à la fin des années 1990. De la fin des années 2000 jusqu'à aujourd'hui, un nombre conséquent de travaux ont été consacrés à ce sujet. Notre objectif ici est de traiter cette question dans le cas où la variable d'intérêt est catégorielle et dans un cadre général regroupant la modélisation logit dichotomique et polytomique. Les solutions discutées dans cet article consistent à passer par une traduction du coefficient logit sous forme de probabilités. Après avoir rappelé les données du problème et les différents registres de solutions proposées, nous étudions le comportement de deux méthodes de traduction d'un coefficient logit, l'écart expérimental et l'écart pur, du point de vue de la comparaison entre plusieurs groupes des résultats d'une modélisation logit.

Keywords
Average marginal effect, experimental deviation, logit modeling, logistic regression, pure deviation

Mots clés
Écart expérimental, écart pur, effet marginal moyen, modèle logit, régression logistique

L'usage des méthodes de régression s'est fortement développé dans le domaine des sciences sociales au cours de ces quarante dernières années. La modélisation d'une variable dépendante catégorielle a d'abord été réalisée par le biais d'une régression linéaire, puis progressivement remplacée par la régression dite logistique, qui consiste à modéliser le logit de la probabilité de la variable dépendante[1]. Le modèle est dit dichotomique lorsque la variable dépendante est binaire, et polytomique lorsqu'elle possède plus de deux modalités.

La généralisation de l'utilisation du modèle logit s'est accompagnée d'un ensemble de pratiques de modélisation. On compte, parmi celles-ci, la comparaison des effets d'une variable d'un même modèle réalisé sur plusieurs groupes, soit à une même date entre deux populations différentes, soit sur une même population à deux dates différentes. On cherche ainsi dans le premier cas à comparer l'effet d'une même variable sur deux populations différentes, et dans le second cas à étudier l'évolution dans le temps de l'effet d'une variable donnée sur une même population.

La façon la plus courante de répondre à cette question consiste à réaliser un même modèle sur les deux populations, puis à procéder à une comparaison du coefficient logit de la variable d'intérêt estimé sur chacune d'elle. On réalise ainsi par exemple un modèle visant à étudier l'effet du diplôme sur la probabilité de devenir cadre à 30 ans, en y incluant bien entendu d'autres variables de contrôle, qu'on va appliquer sur l'enquête

Emploi de l'Insee en 1990 et en 2010, permettant ainsi d'étudier l'évolution dans le temps du poids du diplôme sur la probabilité d'accéder à une position de cadre[2].

Or, cette pratique, très courante dans le cas de la modélisation linéaire classique, soulève de sérieuses objections méthodologiques dans le cas de la modélisation logit[3]. Ce point a été mis en évidence pour la première fois dans la littérature sociologique par Allison à la fin des années 1990 (Allison, 1999). De la fin des années 2000 jusqu'à aujourd'hui, un nombre conséquent de travaux ont été consacrés à cette question dans la littérature sociologique de langue anglaise. Ces travaux diffèrent selon le type de solutions proposées mais également selon le type de modélisation envisagée, qui vont du modèle logit ou probit dichotomique jusqu'aux modélisations appliquées aux données de panels[4]. Les solutions envisagées relèvent cependant pour l'essentiel du cas où la variable explicative d'intérêt est numérique et dans le cadre d'une modélisation dichotomique.

Cet article s'inscrit dans ce débat avec un double objectif : d'une part discuter les données du problème et le principe des solutions envisagées pour y remédier (points 1, 2 et 3), et d'autre part traiter cette question dans le cas où la variable d'intérêt est catégorielle et dans un cadre général regroupant la modélisation logit dichotomique et polytomique (points 4, 5 et 6), ce qui correspond aux situations les plus courantes rencontrées en sociologie.

Les solutions discutées dans cet article consistent à passer par une traduction du coefficient logit sous forme de probabilités, et s'inscrivent dans la lignée de celle proposée par Long (Long, 2009 ; Long et Mustillo, 2018). Il existe en effet deux façons simples et pertinentes de traduire sous forme de probabilités un coefficient logit lorsque la variable explicative est catégorielle : la première, que nous appelons l'écart expérimental, correspond dans la littérature anglo-saxonne à une des formes de l'effet marginal moyen (*Average Marginal Effect*), la seconde, que nous appelons écart pur, a été proposée par Laurent Toulemon (Leridon et Toulemon, 1997; Deauvieau, 2010 et 2011). Nous nous proposons d'observer le comportement de ces deux méthodes de traduction lors de la comparaison entre plusieurs groupes des résultats d'une modélisation logit. Après avoir rappelé les données du problème et les différents registres de solutions proposées, nous exposons les principes des deux méthodes de traduction d'un coefficient logit, pour ensuite les confronter à la question de la comparaison entre groupes.

La nature du problème

Partons d'un exemple simulé pour saisir la nature du problème. Un modèle logit est réalisé pour expliquer une variable dichotomique – le fait de devenir cadre (*versus* non cadre) – par le niveau de diplôme (en deux modalités : égal ou supérieur au bac *versus* inférieur au bac) et la localisation géographique qui sert ici de variable de contrôle (également en deux modalités : grande agglomération *versus* petite agglomération). Pour comparer l'effet du diplôme sur la probabilité de devenir cadre chez les hommes et chez les femmes, le même modèle est appliqué sur chacune des deux populations. Le résultat est le suivant : le coefficient logit lié à la variable diplôme vaut 2,35 chez les hommes et 1,4 chez les femmes (tableau 1). Autrement dit, dans les deux cas être diplômé de l'enseignement supérieur fait augmenter la probabilité de devenir cadre, mais cet effet

Tableau 1. Modèle logistique et modèle linéaire

	Hommes		Femmes	
	modèle logistique	modèle linéaire	modèle logistique	modèle linéaire
Intercept	-2.31	0.15	-1.18	0.25
Niveau études égal ou supérieur au bac	2.35	0.30	1.40	0.30
Grande agglomération	4.17	0.63	1.38	0.30

est nettement plus important pour les hommes que pour les femmes. En appliquant sur les mêmes données un modèle de régression linéaire[5], le résultat est sensiblement différent : le coefficient linéaire lié à la variable diplôme, qui vaut ici 0,3, est identique pour les hommes et pour les femmes (Tableau 1).

Les données de cet exemple simulé ont été construites de la façon suivante :

1/ le nombre d'hommes et de femmes est équivalent.
2/ L'effet du diplôme sur la probabilité de devenir cadre est le même au plan général pour les hommes et pour les femmes. Ainsi la probabilité de devenir cadre, chez les hommes comme chez les femmes, est de 0,4 pour celles et ceux qui ont un niveau d'études inférieur au bac et de 0,7 pour les autres.
3/ Chez les hommes comme chez les femmes, la répartition du niveau du diplôme est indépendante de la localisation géographique. Ainsi, chez les hommes, on trouve, dans les petites agglomérations comme dans les grandes agglomérations, 60 % d'individus ayant un niveau d'études inférieur au bac et 40 % un niveau d'études supérieur au bac. Chez les femmes, ces proportions sont respectivement de 48 % et 52 %.
4/ En revanche, l'effet de la localisation géographique sur la probabilité de devenir cadre est plus élevé chez les hommes que chez les femmes. Ainsi, chez les hommes la probabilité de devenir cadre passe de 0,30 lorsque l'individu habite dans une petite agglomération à 0,92 dans le cas d'une grande agglomération ; et chez les femmes, la probabilité de devenir cadre passe entre ces deux situations de 0,39 à 0,69.

Résumons. Les variables localisation géographique et diplôme ne sont pas liées entre elles chez les hommes comme chez les femmes, et l'effet du diplôme sur la probabilité de devenir cadre est le même dans les deux cas. Dans ces conditions, pourquoi le coefficient logit liant la variable diplôme à la probabilité de devenir cadre est-il plus élevé pour les hommes que pour les femmes, alors que le coefficient linéaire, lui, est strictement équivalent ?

La différence de comportement entre modèle linéaire et modèle logistique tient ici à une propriété du modèle logistique. Dans le cas linéaire, la valeur du coefficient est indépendante de l'hétérogénéité non observée du modèle, alors que ce n'est pas le cas en modélisation logistique. Dans notre exemple, l'hétérogénéité globale du modèle est plus forte chez les femmes que chez les hommes car si l'effet du diplôme est le même dans les

deux populations, en revanche la variable localisation géographique est nettement plus corrélée à la probabilité de devenir cadre chez les hommes. Le modèle est donc globalement mieux ajusté aux données pour les hommes que pour les femmes, et c'est cela qui produit un coefficient logit lié à la variable diplôme plus élevé chez les hommes.

Pour comprendre cet effet de l'hétérogénéité non observée sur le coefficient logit, revenons au principe de cette modélisation. Une façon courante de présenter l'approche théorique de la modélisation logit consiste à passer par une variable latente[6]. Soit Y la variable à expliquer observée qui prend les valeurs 0 ou 1. On postule l'existence d'une variable latente Z, inobservée et continue, qui est liée à Y par la relation suivante :

$$\text{Si Z} > 0, \text{ alors Y} = 1$$

$$\text{Si Z} \leq 0, \text{ alors Y} = 0.$$

On cherche à expliquer Z par une variable X_1, et on pose que Z est liée à X_1 par le biais d'un modèle linéaire :

$$Z = V_0 + V_1 X_1 + \mu$$

Pour un individu i, on a donc :

$$Z_i = V_0 + V_1 X_{1i} + \mu_i$$

La probabilité que l'individu i soit dans l'état $Y = 1$ est égale à :

$$P_i = P(Y_i = 1) = P(Z_i > 0) = P(V_0 + V_1 X_{1_i} > -\mu_i) \quad (1)$$

Notons F la fonction de répartition de $-\mu$, définie par $F(X) = P(-\mu < X)$.

Si on admet que $-\mu_i$ suit une loi logistique, alors on a :

$$F(X) = \frac{1}{1 + e^{(-X)}} \quad (2)$$

On obtient donc:

$$P_i = \frac{1}{1 + e^{(-(V_0 + V_1 X_{1_i}))}} \quad (3)$$

Ou encore, en réarrangeant les termes :

$$Ln \frac{P_i}{1 - P_i} = V_0 + V_1 X_{1_i} \quad (4)$$

L'équation (4) donne la présentation courante du modèle logit. Cependant, cette écriture suggère que l'estimation du coefficient V_1 est indépendante du résidu μ. Or, ce n'est pas le cas. Reprenons pour cela l'étape du raisonnement qui porte sur la nature de μ. Nous avons postulé que μ suit une loi logistique qui possède une variance fixe et à peu près égale à 3,29 (sa valeur exacte est $\frac{\pi^2}{3}$). Cette variance est arbitraire, et rien n'indique que la variance réelle de μ sera effectivement égale à 3,29. C'est pourquoi il convient de passer par une étape intermédiaire.

On pose :

$$\mu = \sigma \times \epsilon$$

Avec ϵ qui suit effectivement une loi logistique avec une variance fixe de 3,29 et σ qui est un coefficient strictement positif permettant d'ajuster à la hausse ou à la baisse ϵ afin de retrouver la vraie valeur du résidu μ. Il ne reste plus ensuite qu'à diviser les termes de l'équation (1) par σ pour retrouver un résidu qui suit une loi logistique. On obtient l'expression suivante:

$$P_i = P\left(\frac{V_0}{\sigma} + \frac{V_1}{\sigma} X_{1_i} > -\frac{\mu_i}{\sigma}\right) = P\left(\frac{V_0}{\sigma} + \frac{V_1}{\sigma} X_{1_i} > -\epsilon_i\right) \quad (5)$$

Avec par construction ϵ qui suit une loi logistique. Un modèle logit n'estime donc jamais les vrais coefficients V_0 et V_1 mais des coefficients B_0 et B_1 qui sont égaux à :

$$B_0 = \frac{V_0}{\sigma} \quad \text{et} \quad B_1 = \frac{V_1}{\sigma}$$

En d'autres termes, le « vrai » coefficient logit du modèle n'est estimé qu'à un facteur inconnu σ près, dont la valeur dépend du résidu μ, donc de l'hétérogénéité globale du modèle. C'est cette propriété qui peut rendre délicate la comparaison directe des coefficients logits entre un même modèle réalisé sur des populations différentes. En effet, une différence entre deux coefficients logits estimés B_1 et B_2, correspondant à l'effet d'une même variable mais calculé sur deux populations différentes P1 et P2, peut venir soit d'une différence d'effet de la variable explicative associée– donc d'une différence entre V_1 et V_2 ; soit d'une différence entre σ_1 et σ_2 – donc d'une différence entre les deux modèles du point de vue de l'hétérogénéité non observée ; soit bien sûr de toute combinaison entre les deux phénomènes.

Concrètement, si un coefficient logit d'un même modèle est plus élevé dans une population 1 que dans une population 2, cela peut s'expliquer : soit par le fait que l'hétérogénéité non observée par le modèle est moins forte pour la population 1 que pour la population 2, soit par le fait que l'effet de la variable liée au coefficient logit est plus fort pour la population 1 que pour la population 2, soit par toutes combinaisons des deux explications. C'est cette propriété du coefficient logit qui fait problème lors de la comparaison des effets d'une variable entre différents groupes.

Les différents registres de solutions

Le fait qu'un coefficient logit n'est identifié qu'à un facteur d'échelle près est une propriété connue depuis le milieu des années 1980 dans la littérature spécialisée en statistique ou en économétrie. En revanche, le fait que cette propriété peut rendre problématique la comparaison directe des coefficients logits entre plusieurs populations n'est clairement signalé par Allison que quinze ans plus tard dans la littérature méthodologique en sociologie quantitative (Allison, 1999).

Les précurseurs de cette réflexion - Allison et Long - étaient en train d'étudier les différences sexuées en matière d'avancement des carrières universitaires aux Etats-Unis. En réalisant deux modèles séparés pour les hommes et les femmes, ces chercheurs

mettent en évidence que l'effet de la productivité scientifique sur la probabilité d'obtenir une titularisation comme professeur semble plus important pour les hommes que pour les femmes, puisque le coefficient logit lié à la variable du nombre d'articles publiés est plus élevé pour les hommes (Long et al., 1993). Le résultat retient l'attention : il suggère que le fonctionnement du monde académique est sexué et que l'activité de publications des femmes est moins reconnue en matière de progression de carrière académique que celle des hommes.

Étant donné l'importance sociologique et politique du constat, on comprend bien qu'il convient de s'assurer de la solidité du résultat statistique. Or, et c'est bien là le point de départ de l'identification du problème, deux explications de cette différence entre les coefficients logits sont possibles et peuvent se cumuler : soit le coefficient logit est plus élevé pour les hommes parce qu'effectivement le rendement à la publication en terme de carrière académique est sexué ; soit il est plus élevé parce que l'hétérogénéité du modèle est plus élevée pour les femmes, par exemple parce que leurs carrières académiques sont moins linéaires que celles des hommes. Les conclusions tirées ne sont pas nécessairement de même nature selon le cas de figure envisagé. Afin de sortir de cette indétermination dans la lecture du résultat statistique, trois grands registres de solutions ont été envisagés.

La première solution consiste à revenir à un modèle de régression linéaire directement appliqué sur la variable à expliquer dichotomique, qu'on appellera donc un modèle linéaire de probabilités. Effectivement, le modèle linéaire est certainement moins adapté pour modéliser des probabilités[7], mais les coefficients linéaires ne sont pas affectés par l'hétérogénéité non observée, et sont donc comparables entre des populations différentes et donc facilement interprétables. Le modèle linéaire de probabilités a, pour ces raisons, fait un grand retour dans la littérature sociologique depuis une dizaine d'années[8]. Notons toutefois que le modèle linéaire n'est véritablement applicable que dans le cas dichotomique. La modélisation logit continue donc d'être la seule alternative dans le cas polytomique.

Le deuxième registre de solutions consiste à rester dans le cadre de la modélisation logit mais en réalisant une transformation de la variable dépendante Y et/ou une forme ou une autre de standardisation du coefficient logit. Allison est le premier à s'être engagé dans cette voie (Allison, 1999). Ces transformations peuvent permettre, sous certaines conditions ou hypothèses parfois restrictives, de comparer les coefficients logits entre des populations différentes en cherchant à « corriger » l'effet de l'hétérogénéité non observée (Allison, 1999 ; Williams, 2009).

Le troisième registre de solution consiste également à conserver une modélisation logit mais en réalisant la comparaison des résultats entre les différentes populations non pas à partir des coefficients mais à partir des probabilités estimées par le modèle (Long, 2009 ; Long et Mustillo, 2018). Cette voie, initialement proposée par Long, repose sur le fait que, contrairement au coefficient logit, les probabilités individuelles estimées par le modèle sont insensibles à l'hétérogénéité non observée du modèle. Revenons aux équations précédentes pour comprendre cette particularité.

L'expression du modèle logit sous forme de probabilités est la suivante :

$$P_i = P(Y_i = 1) = P(Z_i > 0) = P(V_0 + V_1 X_{1_i} > -\mu_i) \quad (1)$$

Et nous avons posé au-dessus :

$$B_0 = \frac{V_0}{\sigma} \quad \text{et} \quad B_1 = \frac{V_1}{\sigma}$$

et ϵ qui suit une loi logistique, alors l'expression du modèle logit devient :

$$P_i = P\left(\frac{V_0}{\sigma} + \frac{V_1}{\sigma}X_{1_i} > -\frac{\mu_i}{\sigma}\right) = P\left(\frac{V_0}{\sigma} + \frac{V_1}{\sigma}X_{1_i} > -\epsilon_i\right) \quad (5)$$

Entre les équations (1) et (5) la valeur du coefficient logit a changé, en revanche l'estimation de la probabilité P_i est restée la même dans les deux cas, puisqu'entre ces deux équations nous avons divisé tous les termes de l'inégalité par le coefficient σ. On a donc bien :

$$P_i = P(Y_i = 1) = P(Z_i > 0) = P(V_0 + V_1X_{1_i} > -\mu_i) = P(B_0 + B_1X_{1_i} > -\epsilon_i)$$

Il s'ensuit que les estimations des probabilités individuelles sont bien indépendantes de l'importance du résidu μ et peuvent donc être utilisées dans le cadre de la comparaison des résultats entre des populations différentes.

Le problème soulevé est-il vraiment un problème ?

La question soulevée par Allison à la fin des années 1990 a entrainé ces dix dernières années la production d'une abondante littérature dans le domaine de la méthodologie statistique appliquée aux sciences sociales. Si les auteurs proposent une diversité de solutions, tous s'accordent à considérer que le problème posé est sérieux[9]. Deux articles très récents apportent cependant un son de cloche différent (Buis, 2017 ; Kuha et Mills, 2018). Ces travaux ne remettent pas en cause les propriétés particulières du modèle logit que nous venons d'aborder, mais récuse leur caractère problématique. Buis estime notamment que ces propriétés sont au contraire tout à fait pertinentes dans le cadre d'une modélisation qui porte sur des probabilités (Buis, 2017).

Le raisonnement de ces auteurs débute par une discussion sur le statut de la variable latente dans la modélisation logit. Rappelons-en le principe : derrière une variable catégorielle Y dichotomique (codée en 0/1), que nous observons dans nos données, se trouve une variable continue Z inobservée qui quantifie une propension à adopter le comportement observé. Nous posons ainsi, comme vu au-dessus :

$$\text{Si Z} > 0, \text{ alors Y} = 1$$

$$\text{Si Z} \leq 0, \text{ alors Y} = 0.$$

Il y a deux façons d'envisager le statut de cette variable inobservée Z. Soit, option 1, elle est considérée comme réelle et l'objet de la modélisation est *in fine* de raisonner sur cette variable Z. Soit, option 2, elle est considérée comme purement théorique, comme une idéalité mathématique, et l'objet de la modélisation ne concerne que le comportement réellement observé – saisi par la variable Y – et la probabilité P de sa réalisation : P (Y=1).

Tableau 2. Probabilité de devenir cadre chez les diplômés et les non diplômés

diplôme	non cadre	cadre	total	probabilité	coefficient logit	coefficient linéaire
0	500	300	800	0,38	1,02	0,25
1	300	500	800	0,63		

Tableau 3. Probabilité de devenir cadre chez les diplômés et les non diplômés en fonction du sexe

sexe	diplôme	non cadre	cadre	total	probabilité	coefficient logit	coefficient linéaire
homme	0	200	200	400	0,50	1,10	0,25
homme	1	100	300	400	0,75		
femme	0	300	100	400	0,25	1,10	0,25
femme	1	200	200	400	0,50		

Les auteurs indiquent que l'impact de l'hétérogénéité non observée sur le coefficient logit n'est véritablement un problème que si l'on se place dans l'option 1. Car dans ce cas l'objectif de la modélisation est bien de raisonner sur cette variable Z inobservée directement dans les données et dont on ne connait pas au fond l'unité réelle. En revanche, si l'on se place dans l'option 2, c'est-à-dire exclusivement sur le plan des comportements réellement observés et dans le langage des probabilités, alors le problème signalé n'en est plus un, à partir du moment bien sûr où l'utilisateur est au clair sur les propriétés du modèle logit.

On peut aisément souscrire à l'option 2, qui correspond bien à l'usage des variables catégorielles dans la plupart des cas en sciences sociales, et considérer l'option 1 comme un usage théorique particulier. La question posée devient alors : en adoptant l'option 2, que fait-on réellement lorsqu'on utilise un modèle logit ? Pour y répondre, il faut bien toujours avoir en tête les propriétés du modèle logit, qu'on peut facilement mettre en évidence en les comparant avec celles du modèle linéaire. Observons donc ces propriétés à partir des deux exemples empiriques simulés ci-dessous.

Exemple 1[10]

J'étudie l'effet du diplôme (en 0/1) sur la probabilité de devenir cadre. 1600 individus sont interrogés, 800 diplômés et 800 non diplômés. La probabilité de devenir cadre chez les diplômés est égale à 0,63 et chez les non diplômés à 0,38 (tableau 2). On peut comparer ces deux situations en utilisant un coefficient logit – qui correspond donc à la différence du logit entre les diplômés et les non diplômés – qui vaut ici 1,02[11] ; ou en utilisant un coefficient linéaire – qui correspond à la différence de probabilités entre les diplômés et les non diplômés – qui vaut ici 0,25 (0,63 – 0,38).

J'introduis maintenant la variable sexe dans le raisonnement (tableau 3). On observe alors l'effet du diplôme chez les hommes (800 individus) et chez les femmes (également 800 individus). Chez les hommes, la probabilité de devenir cadre passe de 0,5 pour les

non diplômés à 0,75 pour les diplômés ; et chez les femmes respectivement de 0,25 à 0,50. Le coefficient logit chez les hommes vaut 1,1, comme chez les femmes ; et le coefficient linéaire vaut dans les deux cas 0,25.

On remarquera que le coefficient linéaire ne change pas selon les situations : il vaut 0,25 dans la population globale et dans chacun des sous-ensembles observés, c'est-à-dire chez les hommes et chez les femmes. En revanche, le coefficient logit passe de 1,02 (tableau 2) lorsqu'il est calculé sur la population d'ensemble, hommes et femmes réunis, à 1,1 lorsqu'il est calculé séparément chez les hommes et chez les femmes (tableau 3).

Cette différence entre coefficient logit et linéaire peut se résumer de la façon suivante :

- Un coefficient linéaire est additif, autrement dit si le coefficient vaut A dans deux sous-populations, alors il vaudra également A s'il est calculé sur l'ensemble constitué de la réunion des deux sous-populations.
- Un coefficient logit est sous-additif, autrement dit si le coefficient logit vaut A dans deux sous-populations, alors il vaudra B s'il est calculé sur l'ensemble constitué de la réunion des deux sous-populations, avec $B < A$.

Exemple 2

A partir d'une nouvelle enquête, j'étudie toujours l'effet du diplôme (0/1) sur la probabilité de devenir cadre en comparant les hommes et les femmes, et en contrôlant cet effet avec la variable taille de l'agglomération (toujours en 0/1). Il s'agit là des données de l'exemple simulé présenté en début d'article (tableau 4).

Le coefficient linéaire de l'effet du diplôme vaut 0,3 chez les hommes comme chez les femmes et le coefficient logit vaut 2,27 dans les deux cas. Si l'on se place maintenant dans les sous ensemble des petites et grandes agglomérations, alors les coefficients logits et linéaires seront à chaque fois différents chez les hommes comme chez les femmes (cf. les deux dernières colonnes du tableau 4).

En réalisant un modèle linéaire séparé pour les hommes et pour les femmes et en introduisant comme variable explicative le diplôme et l'agglomération, le coefficient linéaire sera dans les deux cas toujours égal à 0,3. En revanche, le coefficient logit lui ne sera pas égal à 2,27, comme dans le tableau 4 mais variera entre 2,35 pour les hommes et 1,4 pour les femmes, comme nous l'avons vu au début de l'article (tableau 1).

En quoi ces propriétés du modèle logit, comparées à celles du modèle linéaire, sont intéressantes, selon Buis (2017) notamment ? La réponse implique de s'arrêter un instant sur le statut épistémologique que l'on accorde à la notion de probabilité. Buis part d'une conception des probabilités, inscrite dans la continuité de celle proposée par Keynes, qu'il définit comme une façon de quantifier l'incertain. Dans cette perspective, les propriétés du modèle logit que nous venons de décrire deviennent désirables.

Pourquoi ? Revenons sur l'exemple 1 ci-dessus. Nous partons de l'effet du diplôme sur la probabilité de devenir cadre, qui vaut 1,02 en coefficient logit et 0,25 en coefficient linéaire. Lorsque j'introduis dans le modèle la variable sexe, donc, en suivant cette conception de la notion de probabilité, en ajoutant une nouvelle information qui permet de réduire l'incertitude, j'obtiens pour les hommes comme pour les femmes un nouveau coefficient logit qui vaut 1,1, alors que le coefficient linéaire vaut quant à lui toujours

Tableau 4. Probabilité de devenir cadre chez les diplômés et les non diplômés en fonction du sexe et de la taille de l'agglomération

sexe	Agglomération	Diplôme	non cadre	cadre	N	Probabilité	coefficient logit	coefficient linéaire	coefficient logit diplôme	coefficient linéaire diplôme	coefficient logit agglomération	coefficient linéaire agglomération
Homme	0	0	270	30	300	0,10	2,20	0,40	2,27	0,30	3,76	0,60
	0	1	150	150	300	0,50						
	1	0	30	170	200	0,85	1,21	0,10				
	1	1	10	190	200	0,95						
Femme	0	0	180	60	240	0,25	1,27	0,29	2,27	0,30	2,24	0,30
	0	1	110	130	240	0,54						
	1	0	120	140	260	0,54	1,55	0,31				
	1	1	40	220	260	0,85						

0,25. Le contraste logistique entre diplômés et non diplômés augmente ici car j'ai apporté une information supplémentaire – le sexe, qui est une variable liée à la variable à expliquer – qui permet de réduire l'incertitude concernant la probabilité de devenir cadre et qui de ce fait augmente les contrastes entre diplômés et non diplômés.

Il faut donc bien s'entendre sur l'interprétation que l'on donne au changement de la valeur d'un coefficient lorsqu'on introduit une nouvelle variable dans un modèle :

- Dans le cas linéaire, le coefficient lié à une variable donnée A ne changera pas si j'introduis une nouvelle variable liée à la variable à expliquer mais indépendante de A.
- Dans le cas logistique, le coefficient lié à une variable donnée A augmentera nécessairement si j'introduis une nouvelle variable explicative qui est indépendante de A mais qui est liée à la variable à expliquer. Cette nouvelle variable apporte de l'information au modèle et fait donc augmenter mécaniquement le contraste entre les situations indiquées par le coefficient logit lié à la variable A.

Il faut également souligner que lors de la comparaison d'un coefficient entre deux sous-populations :

- Dans le cas linéaire, le coefficient sera le même si l'effet de la variable est linéairement équivalent dans chacune des sous-populations.
- Dans le cas logistique, une différence dans la valeur du coefficient pourra provenir aussi bien d'une différence de l'effet de la variable d'intérêt dans chacune des sous-populations que d'une différence globale de la qualité prédictive du modèle entre ces deux sous-populations. Autrement dit, la valeur d'un effet – donc le coefficient logit - est dépendante du groupe auquel il s'applique.

Si ces caractéristiques sont connues, alors les propriétés du modèle logit dans le cas de la comparaison entre populations sont effectivement tout à fait acceptables. Le problème essentiel, et bien réel à mon sens dans la pratique quotidienne de la recherche en sciences sociales, est que beaucoup de chercheurs utilisent un modèle logit pour comparer l'effet d'une variable entre plusieurs populations – ou pour observer l'évolution d'un coefficient logit entre différents modèles appliqués sur une même population – en ayant implicitement en tête les propriétés du modèle linéaire, et peuvent donc être amenés à se méprendre sur l'interprétation des variations observées de la valeur des coefficients logits.

Reformulons donc la question posée en début d'article : peut-on trouver une méthode permettant de comparer l'effet propre d'une variable catégorielle – indépendamment donc de la qualité globale du modèle – dans deux sous-populations, et cela dans le cas dichotomique comme polytomique ? C'est à l'examen de cette question que sont consacrées les deux sections suivantes.

Deux méthodes de traduction : écart pur et écart expérimental

Le passage par les probabilités estimées est l'une des voies possibles pour comparer les effets d'une variable entre plusieurs populations à partir d'un modèle logit. Il existe cependant différentes façons de traduire sous forme de probabilités un coefficient logit,

qui n'ont pas le même comportement vis-à-vis de la question de l'hétérogénéité non observée, loin s'en faut (Mood, 2010). Autrement dit, si les probabilités individuelles ne sont pas sensibles à l'hétérogénéité non observée, encore faut-il prendre garde à la façon dont on va utiliser ces probabilités individuelles pour traduire l'effet du coefficient logit.

Nous isolons ici deux méthodes de traduction, que nous intitulons « écart expérimental » et « écart pur »[12], car elles remplissent les conditions suivantes :

1/ il s'agit dans les deux cas de méthodes qui sont spécifiquement adaptées au cas des variables explicatives catégorielles.
2/ l'une comme l'autre sont utilisables dans le cadre d'une modélisation dichotomique mais également dans le cadre d'une modélisation polytomique. Ce point est important, car la traduction sous forme de probabilités des coefficients logits permet dans le cas polytomique de raisonner directement sur les probabilités et non plus seulement sur les rapports des probabilités, ce qui constitue l'un des intérêts majeurs de l'opération de traduction sous forme de probabilités des résultats d'une modélisation logit (Deauvieau, 2011).

L'écart expérimental est une première méthode de traduction qui est basée, comme son nom l'indique, sur l'idée d'expérimentation. L'opération consiste à calculer des probabilités ajustées à partir des probabilités individuelles estimées par le modèle. On calcule pour cela des probabilités simulées en assumant la posture « expérimentale » inhérente aux méthodes de régression multiple. Si par exemple la variable explicative d'intérêt est le sexe, alors les probabilités expérimentales pour les hommes et pour les femmes seront calculées de la façon suivante : on calcule d'abord pour tous les individus de l'échantillon leur probabilité de connaître l'événement modélisé en appliquant pour tous le coefficient logit lié à la modalité homme, puis on répète l'opération en appliquant pour tous le coefficient logit lié à la modalité femme. Il suffit ensuite de calculer dans les deux cas de figure la moyenne des probabilités individuelles sur l'ensemble de l'échantillon pour obtenir la probabilité expérimentale pour les hommes et celle pour les femmes.

Ces deux probabilités se lisent de la façon suivante : si tous les individus de l'échantillon étaient des hommes et qu'ils gardaient les mêmes caractéristiques par ailleurs, leur probabilité moyenne de connaître l'événement modélisé serait de X; si tous les individus de l'échantillon étaient des femmes et gardaient les mêmes caractéristiques par ailleurs, leur probabilité moyenne de connaître l'événement modélisé serait de Y. L'écart entre ces deux probabilités mesure donc l'effet net du sexe sur la probabilité de connaître l'événement modélisé.

Une façon alternative de présenter le principe de l'écart expérimental consiste à souligner qu'il s'agit en fait d'une standardisation directe par population type. En effet, le calcul des probabilités expérimentales pour les hommes et les femmes revient à calculer une moyenne pondérée des probabilités individuelles pour les hommes puis pour les femmes en utilisant la même structure dans les deux cas, à savoir la structure de l'échantillon de la population totale, hommes et femmes réunis. La probabilité pour les hommes est donc calculée en faisant « comme si » les hommes avaient une répartition selon l'ensemble des variables explicatives introduites dans le modèle équivalente à

celle de l'ensemble de la population (hommes et femmes réunis), et à répéter la même opération pour les femmes[13].

La méthode de l'écart expérimental fonctionne également dans le cas polytomique[14]. Il suffit pour cela d'utiliser les formules de transition *ad hoc* pour calculer les probabilités expérimentales qui se lisent exactement de la même façon que dans le cas dichotomique. L'intérêt majeur de cette méthode de traduction est qu'elle est très facile à calculer et ne nécessite pas nécessairement l'usage d'un logiciel spécifique. L'inconvénient est que le résultat obtenu en termes d'écart de probabilités ne respecte plus le contraste logistique[15] tel qu'il est indiqué par le coefficient logit du modèle. En d'autres termes, le coefficient logit est utilisé pour réaliser l'expérience et calculer des probabilités ajustées, mais ces probabilités ajustées calculées *ex-post* ne vérifient plus nécessairement le contraste logistique indiqué par le coefficient logit. Nous reviendrons sur ce point au moment de la comparaison des résultats obtenus par les différentes méthodes de traduction.

L'écart pur est une seconde méthode de traduction, proposée par Laurent Toulemon, qui permet de traduire les résultats sous forme de probabilités, et cette fois-ci en respectant la valeur du coefficient logit estimé par le modèle (Leridon, Toulemon, 1997). Le calcul de l'écart pur est équivalent dans son principe au calcul des effectifs théoriques dans le calcul du khi-deux d'un tableau croisé. Il s'agit en effet dans les deux cas de déterminer les effectifs d'un tableau croisé en respectant les marges du tableau de départ et en appliquant une hypothèse sur les données. Dans le cas du calcul des effectifs théoriques réalisés pour déterminer le khi-deux du tableau, l'hypothèse appliquée est celle de l'indépendance entre les deux variables. Dans le cas du calcul de l'écart pur, l'hypothèse appliquée est celle du respect du contraste logistique (issu du coefficient logit du modèle) entre les modalités du tableau croisé. De nouveaux effectifs sont calculés de telle façon que les marges du nouveau tableau de données sont inchangées par rapport au tableau croisé brut de départ et que les contrastes logistiques entre les modalités du nouveau tableau sont les mêmes que ceux estimés par le modèle logit. Ces deux contraintes font qu'il n'existe qu'une seule solution possible au système considéré. Là encore, cette méthode de traduction fonctionne dans le cas dichotomique mais également dans le cas polytomique.

La comparaison des résultats obtenus

L'écart expérimental et l'écart pur sont deux méthodes alternatives qui permettent de calculer des probabilités ajustées pour des variables explicatives catégorielles à partir d'une modélisation logit, dans le cas dichotomique comme dans le cas polytomique. Leur logique respective est différente. L'écart expérimental répond à la question suivante : « si hommes et femmes avaient les mêmes caractéristiques par ailleurs, quelles seraient leur probabilité moyenne de connaître l'événement modélisé ? », alors que l'écart pur répond à la question suivante : « en l'absence d'effets d'autres variables, quelle serait la probabilité pour les hommes et pour les femmes de connaître l'événement modélisé ? ». En pratique et dans les conditions habituelles de l'enquête en sciences sociales, les résultats obtenus par l'une ou l'autre des méthodes sont le plus souvent très proches, mais peuvent cependant parfois diverger.

Tableau 5. Comparaison des probabilités expérimentales et pures

		Probabilités expérimentales	Probabilités pures
Hommes	inf. bac	0.40	0.29
	égal ou sup bac	0.70	0.81
Femmes	inf. bac	0.40	0.38
	égal ou sup bac	0.70	0.72

Dans le cas dichotomique, l'écart expérimental produira un écart de probabilités entre les situations qui sera toujours inférieur ou égal à celui obtenu par un écart pur, tout en étant bien entendu toujours orienté dans le même sens[16]. Autrement dit, l'écart expérimental entre les deux probabilités donne un contraste logistique qui sera inférieur ou égal à celui exprimé par l'écart pur, ce dernier étant quant à lui égal par construction au contraste logistique indiqué par le coefficient logit estimé par le modèle (voir annexe A).

La comparaison des résultats dans le cas polytomique ne donne pas un résultat aussi évident. Cette situation s'explique par le fait qu'un modèle polytomique modélise non pas la variation d'une probabilité comme dans le cas dichotomique mais la variation d'un rapport de probabilités. La monotonie de l'effet d'une variable explicative sur la variable à expliquer porte alors dans ce cas sur le rapport entre deux probabilités, et non sur la probabilité elle-même. En d'autres termes, il est tout à fait envisageable, pour une variable explicative donnée, d'avoir selon l'endroit où l'on se place dans l'échantillon une inversion de sens de l'effet de cette variable du point de vue des probabilités estimées, ce qui peut produire des résultats différents entre les deux modes de traduction (voir annexe B). De ce fait, même si les résultats obtenus sont là aussi relativement proches, il n'existe pas, dans le cas polytomique, de règles fixes de comportement entre les deux méthodes de traduction. La solution la plus évidente consiste à effectuer la traduction selon les deux méthodes et à en comparer les résultats.

Quelle méthode de traduction utiliser dans la comparaison entre groupes ?

Quel est le comportement de ces deux méthodes de traduction du coefficient logit dans le cadre de la comparaison d'un même modèle sur différentes populations ? Traduisons, pour répondre à cette question, le coefficient logit sous forme de probabilités dans l'exemple mobilisé au début de l'article. Le résultat est clair : l'écart expérimental indique un même écart chez les hommes et les femmes entre diplômés et non diplômés, alors que l'écart pur donne un écart entre les probabilités nettement plus important pour les hommes (tableau 5). L'écart expérimental est ainsi insensible aux variations de l'hétérogénéité non observée entre les groupes, alors que l'écart pur - tout comme le coefficient logit - est lui sensible à l'hétérogénéité non observée.

Pour comprendre cette différence de comportement entre les deux méthodes, il convient de revenir à leur principe de construction. L'écart pur est construit à partir de deux éléments : les marges du tableau croisant le niveau du diplôme et le fait de devenir cadre d'une part, et le coefficient logit d'autre part. Or, les marges des tableaux

croisés sont équivalentes par construction chez les hommes et les femmes, alors que le coefficient logit est plus élevé chez les hommes du fait d'une hétérogénéité globale du modèle moins importante que chez les femmes. Donc l'écart pur est mécaniquement plus élevé chez les hommes. Cette méthode de traduction est sensible à l'hétérogénéité non observée, et conserve donc les mêmes propriétés que le coefficient logit.

L'écart expérimental est quant à lui construit à partir des probabilités individuelles estimées par le modèle, qui, elles, ne sont pas sensibles à l'hétérogénéité non observée. Si entre les différents modèles seule l'hétérogénéité non observée diffère, alors l'écart entre les probabilités estimées dans les différentes situations de contraste va rester constant, et les probabilités expérimentales finales – construites comme une moyenne pondérée des probabilités individuelles estimées – seront insensibles elles aussi à l'hétérogénéité du modèle. L'écart expérimental est donc l'une des solutions permettant de comparer entre plusieurs groupes les résultats d'un même modèle logit, si l'on souhaite contrôler les variations de l'effet de la qualité globale du modèle entre les groupes étudiés.

Conclusion

Rassemblons pour conclure les principaux résultats obtenus et les recommandations qu'il convient d'en tirer du point de vue de l'usage de la modélisation logit en sciences sociales. Deux questions peuvent être distinguées : la première concerne l'usage de la traduction sous forme de probabilités dans le cas dichotomique et polytomique indépendamment de la comparaison entre différents groupes ; la seconde concerne la stratégie à adopter pour comparer les résultats entre différents groupes.

S'agissant de la première question, il est important de souligner que les différentes façons de traduire un coefficient logit sous forme de probabilités ne vont pas nécessairement produire des résultats équivalents. Il est donc préférable ici d'examiner les résultats obtenus par différentes méthodes de traduction pour évaluer les divergences dans le cas empirique considéré. Lorsque la variable d'intérêt est catégorielle, dans le cas dichotomique, trois solutions sont envisageables : produire des probabilités expérimentales à partir d'un modèle logit, produire des probabilités pures également à partir d'un modèle logit, et produire des probabilités expérimentales ou pures (ce qui revient ici au même étant donné les propriétés du modèle linaire) à partir d'une régression linéaire. Ces trois méthodes sont légitimes, et tout à fait intéressantes à comparer en termes de résultats obtenus. Dans le cas polytomique, seules les probabilités expérimentales et pures obtenues à partir d'un modèle logit sont pertinentes, car dans les deux cas la somme des probabilités des différentes modalités de la variable dépendante est par construction égale à 1, ce qui n'est pas le cas avec une régression linéaire réalisée sur chacune des modalités de la variable à expliquer. Il est donc préférable de s'en tenir au cadre de la modélisation logit et de comparer les résultats obtenus par les deux méthodes de traduction.

Sur le plan de la mise en œuvre de ces méthodes, plusieurs logiciels proposent aujourd'hui des solutions pour calculer les probabilités expérimentales ou pures. En se tenant aux logiciels libres diffusés en France, Trideux, développé par Philipe Cibois, calcule directement les probabilités expérimentales et pures dans le cas dichotomique

(Cibois, 2010), et Nicolas Robette a développé un package sous R qui donne également les probabilités expérimentales et pures, à partir d'un modèle logit et linéaire dans le cas dichotomique, et d'un modèle logit dans le cas polytomique[17].

S'agissant de la seconde question, la comparaison des résultats d'une même modélisation effectuée sur plusieurs groupes, si l'objectif est bien d'isoler et de comparer l'effet d'une variable donnée entre ces groupes, indépendamment donc des variations de la qualité du modèle, seule la méthode de l'écart expérimental doit être utilisée, puisque l'écart pur produira des probabilités qui seront dépendantes de l'hétérogénéité non observée. L'écart expérimental peut être mis en œuvre dans le cas dichotomique à partir d'un modèle logit ou d'un modèle linéaire, comme signalé au-dessus. Dans le cas polytomique, la comparaison entre les groupes doit être effectuée à partir d'une modélisation logit et exclusivement à partir de l'écart expérimental.

Remerciements

Cette recherche a débuté au début des années 2010 dans le cadre d'un accueil en délégation à l'Ined, au cours duquel j'ai bénéficié d'excellentes conditions de travail dans un environnement scientifique particulièrement stimulant. Depuis ce moment, plusieurs collègues - tout particulièrement Philippe Cibois, Roland Deauvieau, Olivier Godechot, Nicolas Robette, Marion Selz, Laurent Toulemon et Louis-André Vallet - ont bien voulu discuter mes travaux ou m'aider en réalisant des programmes *ad hoc* utilisés dans cet article : je les en remercie vivement. Bien entendu, les analyses développées ici n'engagent que moi.

Déclaration de conflits d'intérêts

Le, la ou les auteur.e.s déclarent n'avoir aucun conflit d'intérêt potentiel pour tout ce qui concerne le déroulement de la recherche, les droits d'auteur et/ou la publication de cet article.

Financement

Le, la ou les auteur.e.s n'ont bénéficié d'aucun soutien financier particulier relatif au déroulement de la recherche, à ses droits d'auteur et/ou à la publication de cet article.

Notes

1. Le logit de la probabilité correspond à l'expression ln(P/1-P), avec P : probabilité de connaitre l'évènement modélisé.
2. Une autre solution pour répondre à la question posée consiste à mélanger les deux populations étudiées en 1990 et 2010 et à réaliser une interaction entre la variable année d'enquête et la variable diplôme sur la probabilité de devenir cadre. Cette solution a cependant pour conséquence ici de confondre les deux populations du point de vue des autres caractéristiques introduites dans le modèle, ce qui peut se révéler être une hypothèse peu réaliste sur le plan sociologique. La seule solution pour relâcher cette hypothèse consiste à mettre en interaction l'année d'enquête avec toutes les variables introduites dans le modèle, ce qui nous ramène alors exactement à la première solution envisagée ici.
3. La comparaison des coefficients logits fait également problème dans le cas où l'on observe l'évolution d'un coefficient logit donné suite à l'ajout de nouvelles variables dans le modèle. Nous ne traiterons pas ici de ce cas de figure. Pour une présentation générale de cette question,

voir : Mood, 2010. Pour une présentation d'une solution spécifiquement adaptée à cette situation, voir : Karlson et al., 2012.

4. On dénombre ainsi plus d'une dizaine d'articles spécifiquement consacrés à ce sujet entre 2009 et 2018. Cette question a notamment été traitée par : Allison (1999) ; Breen et al. (2018) ; Buis (2017) ; Kuha et Mills (2018) ; Long (2007) ; Long et Mustillo (2018) ; Mood (2010) ; Williams (2009). Concernant l'application de cette question aux données de panels, voir : Landerman et al. (2011) ; Mustillo et al. (2012).
5. Il est devenu aujourd'hui peu courant dans la littérature sociologique d'appliquer un modèle de régression linéaire dans le cas d'une variable à expliquer dichotomique. L'échelle linéaire est en effet réputée peu adaptée aux probabilités (Leridon, Toulemon, 1997). Remarquons cependant que la question fait débat, y compris en sociologie (Cibois, 2000), et que l'utilisation d'un modèle linéaire dans ce cas de figure est courante en économie par exemple (Long, 2007[1997] ; Mood, 2010). Indépendamment de la position adoptée sur cette question, il est intéressant de comparer modélisation linéaire et logit pour bien saisir la nature du problème posé.
6. Pour un développement complet sur la question, voir : Allison, 1999 ; Long, 2007 ; Mood, 2010. La présentation de la modélisation logit à partir du concept de variable latente n'a ici qu'une valeur purement théorique et mathématique, sans donc nécessairement considérer que cette variable latente Z renvoie à une réalité sociologique. Cette écriture permet ici de visibiliser la propriété du modèle logit qui nous intéresse ici.
7. Pour une présentation très claire des propriétés des différentes échelles – additives, multiplicatives, logistiques – appliquées aux probabilités, voir : Leridon, Toulemon, 1997. Sur une question proche, on pourra également se reporter au débat, débuté dans la *Revue française de sociologie* au début des années 1980, autour de la manière de comparer les évolutions dans le temps des inégalités scolaires. Voir notamment à ce propos : Combessie, 2011 ; Vallet, 2007.
8. On pourra relire à ce propos et sous un angle nouveau la controverse méthodologique entre Cibois et Aris et Hagenaars concernant l'utilisation d'un modèle linéaire ou logistique publiée dans le *Bulletin de Méthodologie Sociologique* (voir notamment : Cibois, 2000).
9. L'article de synthèse de Mood de 2010 pose le problème dans sa généralité, qui plus est de manière très pédagogique, et conclut en insistant sur le fait que les sociologues doivent modifier leurs usages de la régression logistique. Le message est d'ailleurs très bien résumé dans le titre de l'article : "Logistic Regression : Why We Cannot Do What We Think We Can Do, and What We Can Do About It" (Mood, 2010). Cette position est aujourd'hui majoritaire dans la littérature. L'article de synthèse le plus récent portant sur l'usage de la modélisation logit revient longuement sur le caractère problématique de l'interprétation des coefficients logits lors de la comparaison entre les groupes (Breen et al., 2018).
10. Je reprends ici les données de base proposées par Buis dans son article (Buis, 2017), en modifiant simplement les noms donnés aux variables par souci de cohérence avec les autres exemples développés ici.
11. Pour rappel, le coefficient logit de 1,02 s'obtient par le calcul suivant : ln (0,63/0,37) – ln (0,38/0,62).
12. L'écart expérimental est utilisé depuis longtemps dans le cas de la modélisation logit polytomique, et correspond à ce que l'on appelle généralement *Average marginal effect*. L'écart pur a été défini par Laurent Toulemon dans le cas du modèle logit dichotomique (Leridon, Toulemon, 1997). Pour une discussion sur l'intérêt de la traduction d'un coefficient logit sous forme de probabilités, voir : Deauvieau, 2010, 2011 ; Selz, 2011.

13. Pour une discussion sur les différentes façons d'éliminer les effets de structure, voir Toulemon, 1992.
14. Cette méthode est présentée notamment dans : Asfa Essafi, 2003. Pour un exemple d'utilisation de l'écart expérimental dans le cas polytomique, voir : Deauvieau, Dumoulin, 2010.
15. On entend ici par contraste logistique l'écart de logit entre les deux modalités d'une variable explicative donnée. Il correspond donc ici au coefficient logit estimé par le modèle.
16. Dans le cas dichotomique, si l'écart pur et l'écart expérimental sont calculés à partir d'une régression linéaire, alors les écarts obtenus seront identiques, puisque dans ce cadre, par construction, les écarts entre les probabilités individuelles estimées par le modèle seront les mêmes quel que soit l'endroit où l'on se place dans l'échantillon. On retrouve ici la propriété d'additivité du modèle linéaire évoquée ci-dessus.
17. Il s'agit de la fonction « translate.logit » du package « GDAtools ».

Références

Allison PD (1999) Comparing Logit and Probit Coefficients Across Groups. *Sociological Methods and Research* (28)3 : 186-208.

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

Breen R, Karlson KB and Holm A (2018) Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models. *Annual Review of Sociology* 44 : 39-54.

Buis ML (2017) Logistic Regression : When Can We Do What We Think We Can Do? *Working paper*, May 29th 2017, University of Konstanz, Department of History and Sociology. Available at: http://www.maartenbuis.nl/wp/odds_ratio_3.1.pdf

Cibois P (2010) Trideux Software Integrates Jérôme Deauvieau's "How to Translate a Logit Model into Probabilities". *Bulletin de méthodologie sociologique* 105 : 53-60.

Cibois P (2000) Observation et modèle linéaire ou logistique : réponse à Aris et Hagenaars. *Bulletin de méthodologie sociologique* 67 : 54-64.

Combessie JC (2011) Analyse critique d'une histoire des traitements statistiques des inégalités de destin. Le cas de l'évolution des chances d'accès à l'enseignement supérieur. *Actes de la recherche en sciences sociales* (3)188 : 4-31.

Deauvieau J (2011) Est-il possible et souhaitable de traduire sous forme de probabilités un coefficient logit ? Réponses aux remarques formulées par Marion Selz à propos de mon article paru dans le BMS en 2010. *Bulletin de méthodologie sociologique* 112 : 32-42.

Deauvieau J (2010) Comment traduire sous forme de probabilités les résultats d'une modélisation logit ? *Bulletin de méthodologie sociologique* 105 : 5-23.

Deauvieau J and Dumoulin C (2010) La mobilité socioprofessionnelle des professions intermédiaires : fluidité, promotion et déclassement. *Économie et Statistique* 431-432 : 57-72.

Karlson KB, Holm A and Breen R (2012) Comparing Regression Coefficients Between Same-sample Nested Models Using Logit and Probit. *Sociological methodology* 42(1) : 286-313.

Kuha J and Mills C (2018) On Group Comparisons With Logistic Regression Models. *Sociological Methods and Research* January 7, online.

Landerman LR, Mustillo SA and Land KC (2011) Modeling Repeated Measures of Dichotomous Data: Testing Whether the Within-Person Trajectory of Change Varies Across Levels of Between-Person Factors. *Social Science Research* 40(5) : 1456-1464.

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

Long JS (2009) Group Comparisons in Logit and Probit Using Predicted Probabilities. *Working paper* draft 2009-06-25.

Long JS (2007) *Regression Models for Categorical and Limited Dependent Variables*, vol. 7 of Advanced Quantitative Techniques in the Social Sciences, Thousand Oaks, CA: Sage.

Long JS, Allison PD and Mc Ginnis R (1993) Rank Advancement in Academic Careers: Sex Differences and the Effects of Productivity. *American Sociological Review* 58 : 703-722.

Long JS and et Mustillo AS (2018) Using Predictions to Compare Groups in Regression Models for Binary Outcomes. *Working paper* draft 2018-03-05.

Mood C (2010) Logistic Regression : Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review* 26(1) : 67-82.

Mustillo AS, Landerman LR and Land KC (2012) Modeling Longitudinal Count Data: Testing For Group Differences in Growth Trajectories Using Average Marginal Effects. *Sociological Methods and Research* 41(3) : 467-487.

Selz M (2011) Pourquoi traduire sous forme de probabilités les résultats d'une modélisation logit ? : Réaction à J. Deauvieau (BMS, 2010). *Bulletin de méthodologie sociologique* 112 : 32-42.

Toulemon L (1992) Population-type et autres méthodes de standardisation : application à la mesure du recours à l'avortement selon la PCS. *Population* 1 : 192-204.

Vallet LA (2007) Sur l'origine, les bonnes raisons de l'usage, et la fécondité de l'odds ratio. *Courrier des statistiques* 121-122 : 59-65.

Williams R (2009) Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Group. *Sociological Methods and Research* 37(4) : 531-559.

Annexe A. Comparaison des résultats entre écart expérimental et écart pur dans le cas dichotomique

Nous cherchons ici à démontrer la propriété suivante

L'écart en probabilités (ou le contraste logistique qui en découle) entre deux modalités d'une variable catégorielle sera toujours inférieur ou égal dans le cas d'un écart expérimental par rapport au cas de l'écart pur.

A chaque probabilité P_0 , on associe une probabilité P_1 par la relation suivante :

$$Logit(P_1) = Logit\ (P_0) + K$$

Avec l'objet Logit défini par : $Ln\ \left(\frac{P}{1-P}\right)$

P_0 et P_1 sont deux probabilités de connaitre l'évènement modélisé, et K est un réel non nul négatif ou positif qui correspond au coefficient logit. K correspond donc au contraste logistique estimé par le modèle logit entre deux modalités d'une variable explicative introduite dans le modèle.

Si on considère une suite de valeurs P_0 quelconques distinctes de $]0;1[$ et la suite des valeurs P_1 associées, on a constaté empiriquement que :

$$|Logit\ (moyenne\ des\ P_1) - Logit\ (moyenne\ des\ P_0)| \ \leq |K|$$

Le membre de gauche de l'inégalité correspond au contraste logistique calculé à partir des résultats de l'écart expérimental et le membre de droite correspond au contraste logistique calculé à partir des résultats de l'écart pur (celui-ci étant par construction égal à K).

Démontrons ce résultat en toute généralité pour K > 0

Brève étude de la fonction Logit : cette fonction est définie sur $]0;1[$ par :

$$Logit\ (X) = \ln\left(\frac{X}{1-X}\right) = \ln X - \ln(1-X)$$

Elle est dérivable sur $]0\ ;\ 1[$ et :

$$Logit\ (X) = \ \frac{1}{X} - \ \frac{-1}{1-X} = \ \frac{1}{X(1-X)}$$

Cette dérivée est strictement positive sur $]0;1[$, Logit tend vers $-\infty$ en 0^+ et vers $+\infty$ en 1.

Logit est une bijection strictement croissante de $]0;1[$ vers R.

Expression de P_1 en fonction de P_0 : pour simplifier l'écriture, nous remplaçons P_0 par X et P_1 par Y.

Pour tout X de $]0;1[$:

$$\ln\left(\frac{Y}{1-Y}\right) = \ln\left(\frac{X}{1-X}\right) + K \tag{1}$$

Prenons les opposés, sachant que $-\ln a = \ln\frac{1}{a}$

$$\Leftrightarrow \ln\left(\frac{1-Y}{Y}\right) = \ln\left(\frac{1}{Y} - 1\right) = \ln\left(\frac{1-X}{X}\right) - K$$

Passons à l'exponentielle :

$$\Leftrightarrow \frac{1}{Y} - 1 \ = \ e^{-K}\left(\frac{1-X}{X}\right) \Leftrightarrow \frac{1}{Y} \ = \ \frac{e^{-X}\ (1-X) + X}{X} \ = \ \frac{(e^K - 1)X + 1}{e^K\ .\ X}$$

Finalement :

$$Y = \ \frac{e^K\ .\ X}{(e^K - 1)X + 1}$$

Remarquons que comme Logit est strictement croissante sur $]0;1[$ et que :

$$Logit\ (Y) > Logit\ (X),\ \text{on a toujours :}\ Y \ > X \tag{*}$$

Étudions à présent la fonction f définie sur $]0;1[$ par :

$$f(X) = \frac{e^K \,.\, X}{(e^K - 1)X + 1}$$

C'est une fonction homographique, du type :

$$X \rightarrow \frac{aX + b}{cX + d}$$

$$\text{Avec} c \neq 0 \text{ et } ad - bc \neq 0$$

D'autre part, le dénominateur est strictement positif sur $]0; 1[$ car :

$$e^K - 1 > 0$$

Cette fonction est donc définie, continue, infiniment dérivable sur $]0; 1[$

$$f'(X) = \frac{e^K}{[(e^K - 1)X + 1]^2}$$

Et :

$$f''(X) = \frac{-2e^X\ (e^K - 1)}{[(e^K - 1)X + 1]^3}$$

Sur $]0; 1[$, on a : $f'(X) > 0$ et $f''(X) < 0$

La fonction f est donc strictement croissante et surtout strictement concave sur $]0; 1[$

Rappelons l'inégalité de Jensen :

Soit f une fonction réelle strictement concave sur un intervalle I, soit une suite $(x_1, x_2 \ldots x_n)$ de n réels distincts de I, et une suite $(a_1, a_2 \ldots a_n)$ de n réels strictement positifs de somme 1, alors :

$$f\left(\sum_{i=1}^{n} a_i X_i\right) \geq \sum_{i=1}^{n} a_i f(X_i)$$

Appliquons ce résultat à la fonction f définie pour une suite de n réels de $]0; 1[$.

On notera m la moyenne des x_i et M la moyenne des y_i . L'inégalité de Jensen (ici tous les coefficients a_i sont égaux à $\frac{1}{n}$) donne :

$$f(m) \geq M$$

Mais d'après (*), pour tout i :

$$y_i \geq x_i$$

Et donc :

$$M \geq m$$

La fonction Logit étant strictement croissante sur $]0; 1[$, on a :

$$m \leq M \leq f(m) \Rightarrow \ Logit\ (m) \leq Logit\ (M) \leq Logit\ (f(m))$$

Mais par définition de *f*,

$$Logit\ (f(m)) - \ Logit\ (m) = K$$

Et donc :

$$0 \leq Logit\ (M) - Logit\ (m) \leq K$$

L'inégalité est donc démontrée dans le cas où K > 0

Pour K < 0, on a : $Logit\ (P_0) = Logit\ (P_1) - K$.

En posant : $K' = -K$, on obtient :

$$0 \ \leq Logit\ (m) - Logit\ (M) \leq \ -K$$

Et donc :

$$|Logit\ (M) - Logit\ (m)| \leq |K|$$

Nous concluons de ce développement que l'écart expérimental sera toujours inférieur ou égal à l'écart pur dans le cas d'une modélisation logit dichotomique.

Annexe B. Comparaison des résultats entre écart expérimental et écart pur dans le cas polytomique

Le résultat démontré dans le cas dichotomique est-il généralisable au cas polytomique ? Pour répondre à cette question, nous partirons du lien existant entre une modélisation logit dichotomique et une modélisation logit polytomique. Les résultats d'un modèle polytomique réalisé sur une variable à expliquer comportant N modalités correspondent aux résultats obtenus à partir des N-1 modèles dichotomiques réalisés sur les populations correspondantes.

Précisons cette proposition à partir de l'exemple suivant :

Sur un échantillon d'individus professions intermédiaires à la date T, nous cherchons à modéliser la probabilité de devenir cadre, employé ou ouvrier ou bien rester profession intermédiaire à la date T1. Nous réalisons pour cela un modèle logit polytomique qui consiste à estimer deux jeux de coefficients qui correspondent par exemple aux deux situations suivantes (en mettant ici la modalité « rester profession intermédiaire » en référence dans la variable à expliquer) :

- Le premier jeu de coefficients correspond à la modélisation de l'évènement « devenir cadre plutôt que rester profession intermédiaire ».
- Le second jeu de coefficients correspond à la modélisation de l'évènement « devenir employé ou ouvrier plutôt que rester profession intermédiaire ».

Le premier jeu de coefficients est strictement équivalent au résultat d'une modélisation logit dichotomique réalisée sur l'échantillon des professions intermédiaires

à la date T devenus cadres ou étant restées professions intermédiaires à la date T1 ; le second jeu de coefficients est strictement équivalent au résultat d'une modélisation logit dichotomique réalisée sur l'échantillon des professions intermédiaires à la date T devenus employés/ouvriers ou étant restées professions intermédiaires à la date T1.

Ce lien entre modélisation dichotomique et polytomique nous permet de proposer le point de départ suivant pour la démonstration :

Soit une modélisation logit polytomique réalisée sur une variable à trois modalités (notées 1, 2, 3). On s'intéresse aux résultats de cette modélisation pour une variable explicative à deux modalités (notées modalité A et modalité B). On obtient donc deux séries de probabilités obtenues à partir du calcul de l'écart pur et deux séries de probabilités obtenues à partir du calcul de l'écart expérimental, notées respectivement :

(P_1 ; P_2 ; P_3) : probabilités estimées pour la modalité A de la variable explicative obtenues par écart pur.
(P'_1 ; P'_2 ; P'_3) : probabilités estimées pour la modalité B de la variable explicative obtenues par écart pur.
(Q_1 ; Q_2 ; Q_3) : probabilités estimées pour la modalité A de la variable explicative obtenues par écart expérimental.
(Q'_1 ; Q'_2 ; Q'_3) : probabilités estimées pour la modalité B de la variable explicative obtenues par écart expérimental.

Propriétés de départ

On a 4 triplés de probabilités vérifiant l'axiome de probabilités totales,

(P_1 ; P_2 ; P_3), (P'_1 ; P'_2 ; P'_3), (Q_1 ; Q_2 ; Q_3), (Q'_1 ; Q'_2 ; Q'_3), 4 constantes réelles a_1, a_2, b_1, b_2.

On a les égalités, inégalités et propriétés suivantes (qui découlent du cas dichotomique étudié dans l'annexe A) :

- $P_1 + P_2 + P_3 = P'_1 + P'_2 + P'_3 = Q_1 + Q_2 + Q_3 = Q'_1 + Q'_2 + Q'_3 = 1$
- $\ln\frac{P_1}{P_3} - \ln\frac{P'_1}{P'_3} = a_1, \ln\frac{P_2}{P_3} - \ln\frac{P'_2}{P'_3} = a_2, \ln\frac{Q_1}{Q_3} - \ln\frac{Q'_1}{Q'_3} = b_1, \ln\frac{Q_2}{Q_3} - \ln\frac{Q'_2}{Q'_3} = b_2$
- a_1 et b_1 d'une part, a_2 et b_2 d'autre part sont de mêmes signes.
- $|a_1| > |b_1|$ *et* $|a_2| > |b_2|$
- $\frac{P_1}{P_3 + P_1} - \frac{P'_1}{P'_3 + P'_1}$ est de même signe que $\frac{Q_1}{Q_3 + Q_1} - \frac{Q'_1}{Q'_3 + Q'_1}$
- $\frac{P_2}{P_3 + P_2} - \frac{P'_2}{P'_3 + P'_2}$ est de même signe que $\frac{Q_2}{Q_3 + Q_2} - \frac{Q'_2}{Q'_3 + Q'_2}$

On cherche à démontrer les deux propriétés suivantes

Propriété 1

- $P_1 - P'_1$ est de même signe que $Q_1 - Q'_1$
- $P_2 - P'_2$ est de même signe que $Q_2 - Q'_2$
- $P_3 - P'_3$ est de même signe que $Q_3 - Q'_3$

Propriété 2

$$|P_1 - P'_1| > |Q_1 - Q'_1| \text{ et } |P_2 - P'_2| > |Q_2 - Q'_2| \text{ et } |P_3 - P'_3| > |Q_3 - Q'_3|$$

Ces deux propriétés ne sont pas vérifiées en toute généralité. On peut en effet trouver un contre-exemple numérique qui vérifie les propriétés de départ mais pas les propriétés 1 et 2 d'arrivée.

$P'_1 = 0.4$ $P'_2 = 0.3$ $P'_3 = 0.3$ $a_1 = \ln 5 a_2 = \ln 2$

Cela donne $P_1 = 0.6897$ $P_2 = 0.2069$ $P_3 = 0.1034$

$Q'_1 = 0.1$ $Q'_2 = 0.4$ $Q'_3 = 0.5$ $b_1 = \ln 1.1$ $b_2 = \ln 1.9$

Cela donne $Q_1 = 0.0803$ $Q_2 = 0.5547$ $Q_3 = 0.3650$

Les conditions imposées sont remplies :

$a_1 > b_1 > 0$ $a_2 > b_2 > 0$

$P_1/(P_1 + P_3) = 0.8616$ $P'_1/(P'_1 + P'_3) = 0.5714$ Différence $= 0.2982$

$Q_1/(Q_1 + Q_3) = 0.1803$ $Q'_1/(Q'_1 + Q'_3) = 0.1667$ Différence $= 0.0136$

$0.2982 > O.O136 > 0$

$P_2/(P_2 + P_3) = 0.6667$ $P'_2/(P'_2 + P'_3) = 0.5000$ Différence $= 0.1667$

$Q_2/(Q_2 + Q_3) = 0.6031$ $Q'_2/(Q'_2 + Q'_3) = 0.4444$ Différence $= 0.1587$

$0.1667 > 0.1587 > 0$

On obtient donc $P_1 > P'_1$ mais $Q_1 < Q'_1$

Ce résultat invalide les propriétés 1 et 2 définies au-dessus. Autrement dit, dans le cas polytomique, il n'existe pas de relations d'inégalités données entre les résultats obtenus avec un écart expérimental et ceux obtenus avec un écart pur. Cependant, en pratique, autrement dit dans le cas de données usuelles en sciences sociales, les résultats obtenus par les deux méthodes ne sont jamais très éloignés.