Bulletin de Méthodologie Sociologique
112 32–42

DOI: 10.1177/0759106311417537
http://bms.sagepub.com

Est-il possible et souhaitable traduire sous forme de probabilites un coefficient logit ? - Réponse aux remarques formulées par Marion Selz à propos de mon article paru dans le *BMS* en 2010

Jérôme Deauvieau
Laboratoire Printemps (UVSQ/CNRS) et Laboratoire de sociologie quantitative (GENES/CREST)

Abstract
Is It Possible and Desirable to Translate a Logit Coefficient into Probabilities – Reply to Comments by Marion Selz on My *BMS* 2010 Article: Marion Selz responded to my *BMS* article on the translation of a logit coefficient into probabilities. I propose here to respond to her objections by showing that the translation is not a problem when one thoroughly understands what takes place. We first show the relationship between linear models and logistic models, and then the relationship between the translation of a logit coefficient and standardization in demography, and finally the sociological interest of such a translation in the general framework of logit modeling.

Résumé
Marion Selz a réagit à mon article paru dans le *BMS* en 2010 portant sur la traduction sous forme de probabilités d'un coefficient logit. Je me propose ici de répondre aux

Corresponding Author:
Email: jerome.deauvieau@uvsq.fr

objections formulées en montrant que la traduction ne pose pas de problème dès lors que l'on comprend bien l'opération réalisée. Nous montrons pour cela d'abord les rapports entre modèle linéaire et modèle logistique, puis les liens entre la traduction d'un coefficient logit et la standardisation en démographie, et enfin l'intérêt sociologique de la traduction dans le cadre général de la modélisation logit.

Keywords
Logit Models, Logistic Regression, Probabilities, Standardization

Mots clés
Modélisation logit, Régression logistique, Probabilités, Standardisation

Marion Selz réagit à mon article paru dans le *BMS* portant sur la traduction sous forme de probabilités d'un coefficient logit (Deauvieau, 2010). J'exposais dans cet article deux façons différentes de traduire sous forme de probabilités les résultats d'une modélisation logit lorsque toutes les variables explicatives sont catégorielles. Marion Selz conteste l'opération de traduction car elle estime qu'il est « incohérent de vouloir transformer en un écart de pourcentages unique un coefficient logit qui a été créé pour harmoniser des écarts de pourcentages se situant à des niveaux de pourcentages différents. De plus, cet écart de pourcentage, lui, n'est pas accompagné d'un écart-type ni d'un seuil de confiance : on perd ainsi la notion de significativité du résultat » (Selz, 2011 : 78). Je me propose ici de répondre à ces deux objections.

Modele logistique et modèle lineaire

Le premier point avancé par Marion Seltz consiste à rappeler qu'un modèle de régression logistique n'est pas linéaire en probabilités. Ainsi, si on trouve un contraste logistique entre hommes et femmes de 0,4 par exemple, cela se traduit par des écarts en probabilités variables selon l'endroit où l'on se place.

On a en effet l'égalité suivant :$1\mathrm{n}\left(\frac{PH}{1-PH}\right) - 1\mathrm{n}\left(\frac{PF}{1-PF}\right) = 0,4$.

Si PF=0,1, alors en appliquant l'égalité du dessus on trouve que PH=0,14.

Si PF=0,5, alors en appliquant l'égalité du dessus on trouve PH=0,6.

Et si PF=0,7 alors en appliquant l'égalité du dessus on trouve PH=0,78.

On montre là l'une des propriétés de la régression logistique. Ce type de régression est bien un modèle de probabilité, au sens où ce qui est modélisé est une probabilité, mais ce modèle est linéaire seulement dans le logit, mais pas dans la probabilité. On le constate de façon évidente lorsqu'on exprime le modèle sous forme de logit (1), ou sous forme de probabilité (2). La première expression est linéaire, la seconde non.

$$1\mathrm{n}\left(\frac{p}{1-p}\right) = \mathrm{B0} + \mathrm{B1X1} \tag{1}$$

$$\mathrm{P} = \frac{1}{1 + \mathrm{EXP}^{(-(B0+B1X1)}} \tag{2}$$

Tableau 1.

Pdépart	ln(P/(1-P))	4P-2	Ecart en valeur absolue	Papproximé avec 4P - 2
0,05	-2,94	-1,80	1,14	0,142
0,1	-2,20	-1,60	0,60	0,168
0,15	-1,73	-1,40	0,33	0,198
0,2	-1,39	-1,20	0,19	0,231
0,25	-1,10	-1,00	0,10	0,269
0,3	-0,85	-0,80	0,05	0,310
0,35	-0,62	-0,60	0,02	0,354
0,4	-0,41	-0,40	0,01	0,401
0,45	-0,20	-0,20	0,00	0,450
0,5	0,00	0,00	0,00	0,500
0,55	0,20	0,20	0,00	0,550
0,6	0,41	0,40	0,01	0,599
0,65	0,62	0,60	0,02	0,646
0,7	0,85	0,80	0,05	0,690
0,75	1,10	1,00	0,10	0,731
0,8	1,39	1,20	0,19	0,769
0,85	1,73	1,40	0,33	0,802
0,9	2,20	1,60	0,60	0,832
0,95	2,94	1,80	1,14	0,858

Il est cependant possible d'aller plus loin en réalisant une approximation linéaire du modèle de régression logistique et en comparant les résultats obtenus. Que se passe-t-il quand j'approxime l'équation (1) par un modèle linéaire en probabilité[1] ? C'est à dire, quand je remplace (1) par un modèle du type : P= B'0 + B1'X1.

Un développement limité d'ordre 1 permet de trouver une approximation correcte au voisinage de P=0,5 avec la forme suivante :

Posons $P = 0{,}5 + \alpha$.

Alors on obtient :

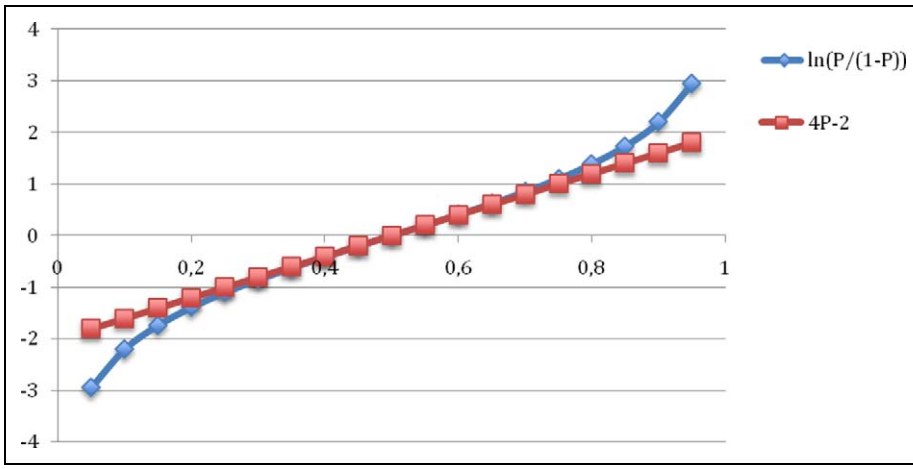
$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{0.5+\alpha}{0.5-\alpha}\right) = \ln\left(\frac{1+2\alpha}{1-2\alpha}\right)$$
$$= \ln(1+2\alpha) - \ln(1-2\alpha) \approx 2\alpha - (-2\alpha) = 4\alpha = 4p - 2$$

On obtient donc, par approximation linéaire, l'égalité

4P-4 = B0 + B1X1 (valable pour P1/4 0,5), soit P= 0,5 + B0 + B1/4.

Donc il est possible d'approximer une régression logistique par une régression linéaire avec l'expression ci-dessus. Il suffit alors de diviser le coefficient logit par 4 pour obtenir un écart en probabilité entre les deux modalités de la variable. Il est, par exemple, possible de dire par approximation qu'un coefficient logit de 0,4 revient à peu près à un écart en probabilités de 0,1 entre les deux modalités considérées.

Quelle est la qualité de cette approximation ? Il suffit de l'observer avec un tableau de valeur (voir Tableau 1 et Graphique 1). L'approximation est parfaite à 0,50 par



Graphique 1.

définition, encore très bonne de 0,5 à 0,25 et de 0,5 à 0,75 puisque nous sommes au maximum à un peu moins de 0,02 points d'erreur (soit souvent très en dessous des intervalles de confiance), et l'approximation devient de plus en plus mauvaise à mesurer qu'on s'éloigne de 0,25 (vers 0) et de 0,75 (vers 1).

Résumons. Il est tout à fait exact de dire que le coefficient logit traduit des écarts en probabilités différents selon le niveau de probabilité où l'on se place. Il est même maintenant possible de préciser que le coefficient logit – entre 0,25 et 0,75 – peut être approximé par un écart en probabilités unique (il suffit de diviser le coefficient logit par 4 pour avoir une très bonne approximation de l'écart en probabilités contenu dans le coefficient logit), et que ce n'est plus le cas entre 0 et 0,25 et entre 0,75 et 1.

Ceci étant établi, je souhaite maintenant discuter le second point de la critique de Marion Selz, à savoir que cette propriété du coefficient logit, sa non linéarité en probabilités, rend « incohérent » le fait de vouloir le traduire sous forme de probabilités ajustées. Pour cela, repartons de la logique même de la traduction d'un coefficient logit.

Traduction D'un Coefficient Logit Et Standardisation

Pour bien comprendre la logique de la traduction d'un coefficient logit sous forme de probabilités, un détour par la standardisation en démographie s'impose. Voici comment Henri Léridon et Laurent Toulemon introduisent l'idée de standardisation :

> Prenons un premier exemple concret, la comparaison des conditions de mortalité en France et au Mexique. En 1991, on a compté 525 000 décès en France, et on estime le nombre de décès au Mexique à 410 000 (OMS, 1994 [Annuaire de statistiques sanitaires mondiales pour 1993. Genèse, OMS]). Rapportés à des populations moyennes de 56,7 et 86,2 millions respectivement, ces décès correspondent à des taux bruts de mortalité de 9,3 et de 4,8 pour mille. Est-ce à dire que l'on meurt plus en France qu'au Mexique ? Oui. Mais la principale raison de la faible mortalité au Mexique est la

Tableau 2.

	homme	femme	ensemble
dipl1	49%	40%	44%
dipl2	23%	25%	24%
dipl3	27%	35%	31%
Total	100%	100%	100%

Tableau 3.

	coefficients	test
intercept	-2,5512	P<0,01
dipl2	0,5605	P<0,01
dipl3	1,1612	P<0,01
Sexe	0,6222	P<0,01

> structure par âge « défavorable » à la mortalité : les Mexicains meurent peu, parce qu'ils sont jeunes. Pour comparer les conditions de mortalité en France et au Mexique, on souhaite donc « éliminer » l'effet de l'âge sur la mortalité au sein de chacune des populations, pour ne garder que ce qui différencie les deux populations. (Leridon et Toulemon, 1997 : 191).

L'objectif de la standardisation est donc bien ici de pouvoir comparer les taux de mortalité « toutes choses égales par ailleurs », indépendamment de la structure par âge. Le principe de la standardisation est fondamentalement le même que celui de la régression logistique. Comment réaliser une standardisation par l'âge ? L'une des solutions est la standardisation directe, dite par population type. Concrètement, on utilise les taux de mortalités par âge pour chaque pays, et on les applique à une population type. Autrement dit, on calcule un nouveau taux moyen de mortalité pour les deux pays en faisant la moyenne pondérée des taux réels par âge de chaque pays, mais sur une population type (donc une structure par âge équivalente) pour les deux pays. Ainsi, la standardisation directe selon l'âge montre que « la mortalité serait plus faible en France qu'au Mexique, si ces deux pays conservaient leurs taux de mortalité par âge mais avaient la structure par âge de la population mondiale (4,1 p. 1000 contre 6,1 p.1000), ou celle de la population Européenne (8,0 contre 11,1 p. 1000) » (Leridon et Toulemon, 1997 : 193).

On remarquera que la différence des taux ajustés varie selon la population type choisie. Henri Leridon et Laurent Toulemon indiquent que l'une des solutions est d'utiliser la structure de l'ensemble des deux sous-populations considérées quand celles-ci font partie d'un ensemble cohérent. Par exemple, si l'on veut standardiser selon l'âge le taux de mortalité des hommes et des femmes d'un pays donné, on peut tout à fait utiliser comme population type l'ensemble des hommes et des femmes de ce pays.

Appliquons maintenant ce raisonnement à la traduction d'un coefficient logit sous forme de probabilités. L'écart expérimental que je présentais dans mon article consiste tout simplement à réaliser une standardisation en utilisant les coefficients du modèle de

Tableau 4.

		effectif de la catégorie	probabilités de devenir cadre	probabilités estimées par le modèle
dipl1	Femme	248	0,08	0,07
	Homme	300	0,12	0,13
dipl2	Femme	158	0,11	0,12
	Homme	143	0,21	0,2
dipl3	Femme	221	0,2	0,2
	Homme	167	0,31	0,32

régression et la structure de l'échantillon. Prenons un exemple pour illustrer la démarche. On chercher à mesurer les différences d'accès à la position de cadre en 5 ans entre hommes et femmes des professions intermédiaires administratives et commerciales. Les femmes dans ces professions sont en moyenne plus diplômées que les hommes (voir Tableau 2), or le diplôme est un facteur important de la mobilité professionnelle. On réalise donc un modèle de régression logistique en incluant comme variables explicatives le sexe et le niveau de diplôme afin de modéliser l'effet propre du sexe (« toutes chose égales par ailleurs », c'est à dire ici indépendamment du niveau de diplôme) sur la probabilité de devenir cadre cinq ans plus tard.

Les résultats de la régression logistique indiquent que lorsque l'on contrôle le niveau de diplôme, le logit de la probabilité de devenir cadre est supérieur de 0,62 pour les hommes par rapport aux femmes (Tableau 3). Pour traduire ce résultat sous la forme de probabilités ajustées en utilisant l'écart expérimental, on réalise les deux opérations suivantes :

- Première opération : calcul des probabilités estimées par le modèle pour chacune des six situations définies par le modèle (colonne 4 du Tableau 4).
- Deuxième opération : pour trouver des probabilités ajustées, on calcule pour les hommes et pour les femmes une probabilité moyenne de devenir cadre en faisant la moyenne des trois probabilités (estimées par le modèle) dans les trois situations et en pondérant avec la répartition moyenne du niveau de diplôme dans l'ensemble de la population (ce qui permet d'annuler l'effet de la répartition différenciée des niveaux de diplôme entre hommes et femmes).

Pour les hommes, le calcul est donc :
PHajustée= ((44*0,13) + (24*0,2) + (31*0,32))/100 = 0,21
Pour les femmes :
PFajustée= ((44*0,08) + (24*0,11) + (31*0,2))/100= 0,12

On obtient donc deux probabilités ajustées qui correspondent à la probabilité moyenne pour les hommes et pour les femmes de connaître une mobilité, s'il avait la même répartition de diplôme que l'ensemble de la population.

Finalement, l'opération de traduction d'un coefficient logit sous forme de probabilités en utilisant l'écart expérimental revient ni plus ni moins à réaliser une standardisation directe telle que nous l'avons décrite au dessus avec le taux de mortalité entre la France

et le Mexique. Le taux de mortalité est remplacé par la probabilité de devenir cadre, la variable sexe remplace la variable pays, et le diplôme remplace l'âge. On calcule la probabilité standardisée pour les hommes en répondant à la question suivante : quelle serait la probabilité d'être mobile pour les hommes s'ils avaient la même répartition par diplôme que l'ensemble constitué des hommes et des femmes ? On réalise la même opération pour les femmes. On obtient les deux probabilités calculées au dessus qui sont bien issues du modèle de régression puisqu'on a utilisé les estimations en probabilités produites par le modèle.

La standardisation des probabilites par une regression logistique est-elle légitime ?

Cette standardisation est-elle légitime ? Elle l'est autant que le fait de standardiser un taux de mortalité selon le pays. Dire qu'on ne peut pas standardiser les probabilités avec les résultats d'une modélisation logit revient à dire qu'on ne peut pas calculer un taux de mortalité par pays car il varie selon l'âge. A la limite, en poussant le raisonnement, on pourrait dire qu'il n'est même pas possible de calculer un taux moyen brut de mobilité pour les hommes . . . puisque ce taux varie selon le niveau de diplôme.

L'opération de traduction d'un coefficient logit consiste donc seulement à calculer des probabilités moyennes et standardisées (ou ajustées pour reprendre la terminologie que j'ai utilisé dans mon article) en utilisant les résultats du modèle de régression. Il s'agit d'un point de vue statistique de répondre à la question suivante : quels seraient les taux moyens de mobilités des hommes et des femmes s'ils avaient les mêmes caractéristiques par ailleurs; c'est à dire, dans notre exemple la même structure par diplôme ? Autrement dit, quels seraient les écarts moyens en probabilités entre les sexes, toutes choses égales par ailleurs. L'intérêt d'utiliser la régression logistique tient bien sûr au fait qu'on peut très facilement ajouter d'autres variables dans le modèle; par exemple, le niveau de diplôme, la CS de départ, etc. Le principe est toujours le même, et revient tout simplement à réaliser une standardisation du taux de mobilité par sexe selon l'ensemble des variables introduites dans le modèle. En faisant cela, aucun principe de la régression n'est violé. La régression logistique est donc une méthode de standardisation efficace et rapide lorsqu'on doit utiliser plusieurs variables dépendantes.

C'est d'ailleurs bien comme cela que Léridon et Toulemon présentent cette méthode en indiquant que « Les régressions tendent à supplanter les méthodes traditionnelles de standardisation en démographie » (Leridon et Toulemon, 1997 : 235). Ils expliquent également que « l'échelle logistique est moins familière que les échelles additives ou multiplicatives, mais on peut la considérer comme l'échelle 'naturelle' des proportions » (Leridon et Toulemon, 1997 : 235). Ils démontrent là ce que l'on a montré dans la première partie du texte et que soulignait Marion Seltz. Cependant, montrer que l'échelle logistique est l'échelle naturelle des proportions n'implique absolument pas selon eux (et moi) l'interdiction de traduire les coefficients logit sous forme de probabilités. Léridon et Toulemon précisent en effet dès la phrase suivante de leur introduction qu'on « verra ensuite comment transformer les paramètres des régressions logistiques en probabilités standardisées » (Leridon et Toulemon, 1997 : 235)[2] Il est donc tout à fait légitime de considérer *à la fois* que l'échelle logistique est l'échelle naturelle des probabilités et qu'il

est possible de transformer les paramètres en probabilités standardisées. C'est pourtant bien la prise en compte de ces deux propositions que conteste au fond Marion Selz lorsqu'elle indique qu'il est « incohérent de vouloir transformer en un écart de pourcentages unique un coefficient logit qui a été créé pour harmoniser des écarts de pourcentages se situant à des niveaux de pourcentages différents » (Selz, 2011: 61). Il y a là, me semble-t-il, une confusion entre deux choses : la nature du coefficient logit, d'une part, et le fait d'utiliser ce coefficient pour calculer des probabilités standardisées d'autre part.

La première proposition porte en effet sur la nature du coefficient logit. Ce coefficient mesurant un contraste entre deux catégories est linéaire sous cette forme dans l'équation de régression, ce qui veut dire que le contraste logistique sera toujours exactement le même quelque soit la situation où l'on se place. C'est là que réside l'intérêt de la régression logistique et ce qui fait du contraste logistique l'échelle naturelle des « proportions ». Le coefficient logit permet de définir par une seule valeur dans l'équation le contraste entre deux situations données et dans le même temps traduire ce contraste par des écarts en probabilités différents, selon l'endroit où l'on se place dans l'échelle des probabilités. L'idée sous-jacente est bien ici le fait qu'un même écart en probabilités ne « vaut » pas la même chose selon que l'on se place aux extrêmes de l'échelle des probabilités ou autour de 0,5. Ainsi, un même coefficient logit correspond à un écart de probabilités entre les deux catégories qui sera faible dans les probabilités proches de 0, puis de plus en plus fort jusqu'à 0,5, puis de plus en plus faible à mesure qu'on s'approche de 1[3]. Au fond, comme l'indique fort justement Léridon et Toulemon, l'échelle logistique est tout à fait intéressante lorsqu'on travaille sur des proportions car « l'échelle logistique est proche d'une échelle multiplicative en p si p est proche de 0 (jusqu'à 20%), et en (1-p) si p est proche de 1 (à partir de 80%), tandis qu'elle est presque additive pour les valeurs moyennes de p (de 30% à 70%) » (Leridon et Toulemon, 1997 : 246).

Ceci étant dit, et c'est là le problème à mon sens dans le raisonnement de Marion Slez, rien dans ce qui vient d'être dit n'empêche de faire la moyenne pondérée par une population type de ces différentes probabilités estimées par le modèle selon les différentes situations pour obtenir des probabilités moyennes ajustées, et donc un écart moyen qui vaut « toutes choses égales par ailleurs ». Cela correspond à la seconde proposition de Leridon et Toulemon et au propos de mon papier dans le *BMS* (Deauvieau, 2010). C'est cela et rien d'autre que réalise une standardisation directe ou un écart expérimental à partir d'une régression logistique.

En conclusion de ce point, il est donc tout à fait clair que le calcul de probabilités ajustées (ou standardisées, ce qui revient au même et est sans doute préférable comme terminologie) à partir de coefficients logit ne pose aucun problème sur le fond, et qu'on peut à la fois respecter la « nature » du coefficient logit et le traduire sous forme de probabilités. L'écart expérimental est une standardisation directe par population type, alors que l'écart pur est une standardisation indirecte (Leridon et Toulemon, 1997 : 252). Chaque méthode de standardisation a ses avantages et ses inconvénients. C'est sans doute ce point qu'il est intéressant de débattre plus que la légitimité du principe de la traduction.

Enfin, dernier point de désaccord, Marion Selz reproche à cette méthode l'absence de test de significativité lors de la traduction sous forme de probabilités, alors que c'est le cas pour le coefficient logit. Il ne me semble pas que cela pose réellement problème.

A partir du moment où un coefficient logit est « significatif », cela veut bien dire que le contraste entre les deux catégories est significativement différent de 0, et donc que l'on peut considérer que l'écart en probabilité ajustée l'est également. Il va de soi qu'il ne faut calculer des probabilités ajustées qu'à partir du moment où le coefficient logit dont elles sont issues est significatif.

L'utilite de la traduction en sociologie

Abordons pour finir la question de l'utilité propre de la traduction. Historiquement, la traduction des résultats d'un modèle logit sous forme de probabilités a été largement utilisée dans le cadre de la modélisation polytomique, car elle permet de découvrir de nouveaux résultats contenus dans le modèle mais pas accessible lors de la lecture du coefficient. En effet, un modèle logit polytomique est utilisé lorsque la variable dépendante comporte plus de deux modalités et consiste en une juxtaposition de logits dichotomiques. Dans ce cas de figure, l'une des modalités de la variable à expliquer sera mise en référence et le logit modélisé correspond alors au log du rapport des probabilités entre une modalité de la variable à expliquer et la modalité de la variable à expliquer mise en référence[4]. Par exemple, si on dispose de trois modalités dans la variable à expliquer et que la modalité 1 est mise en référence, on modélise alors les deux logits suivants :

$$\ln\frac{P2}{P1}, \text{ et } \ln\frac{P3}{P1}$$

et l'estimation comporte deux jeux de coefficients.

Les coefficients logit obtenus nous renseignent sur l'augmentation ou la diminution des deux logits ci dessus, et permet donc de dire si le rapport entre P2/P1 est plus élevé pour telle modalité plutôt que pour telle autre. En revanche, il ne permet pas de dire si P2 est plus élevé, puisque cela dépend du troisième terme P3. On peut tout à fait trouver que le rapport P2/P1 est plus élevé pour les hommes que pour les femmes (coefficient logit positif) sans que P2 soit plus élevé pour les hommes que pour les femmes. Dans ce cas, la traduction sous forme de probabilités est utilisée pour regarder directement les différences entre hommes et femmes sur P1, P2 et P3. Cette méthode permet donc ici d'exprimer des résultats du modèle qui ne sont pas visibles avec les coefficients logit[5].

Dans le cas dichotomique, la traduction ne permet pas de découvrir de nouveaux résultats puisque le sens du coefficient s'interprète directement en termes de probabilités : si un coefficient est positif, cela indique que la probabilité associée est supérieure pour la modalité donnée par rapport à la modalité de référence, et vice versa. J'indiquais dans mon précédent article un intérêt pédagogique de la traduction qui me paraît effectivement indéniable. Mais plus largement, il me semble que la traduction permet surtout de mieux saisir sociologiquement les résultats de la modélisation. Lorsqu'on dit que le contraste logistique entre hommes et femmes est de 0,4, il est intéressant sociologiquement d'ajouter que cela donne un écart de 10 points environ, si la probabilité de départ des femmes est comprise entre 0,25 et 0,75. De la même manière et en suivant la logique, il est intéressant de calculer des probabilités standardisées à partir des coefficients d'une régression logistique. Cela ne pose aucun problème statistique à partir du moment où la démarche est clairement exposée. La substantialisation des résultats sous forme de

probabilités permet d'incarner les différences entre les catégories, et donc de raisonner sociologiquement sur des écarts qui font sens, tout comme le fait la standardisation en démographie.

Plus largement, cette standardisation permet également d'utiliser la modélisation logit en sociologue, c'est à dire en n'oubliant pas que le matériau de base est constitué de tableaux croisés. Une régression logistique de « sociologue » – dans laquelle toutes les variables dépendantes sont mises sous formes de catégories – n'est rien d'autre qu'une modélisation d'un tableau croisé d'une profondeur égale au produit du nombre de modalités introduites dans le modèle[6]. Il est donc tout à fait normal de chercher à exprimer les résultats obtenus sous forme de probabilités en utilisant l'une des méthodes classiques de standardisation. C'est à cette question qu'était consacré mon article précédent dans le *BMS*. Il me paraît maintenant intéressant de poursuivre l'investigation en procédant à une comparaison des résultats obtenus selon les différentes méthodes de standardisation, et à généraliser ces méthodes de standardisation au logit polytomique.

Notes

1. Sur cette question, on pourra consulter, par exemple, Gujarati (2004).
2. Je rappelle d'ailleurs que la méthode de l'écart pur présentée dans mon article a été mise au point par Laurent Toulemon.
3. Sur la question des échelles liées aux proportions, on renvoie au célèbre débat qui a commencé dans la *Revue française de sociologie* sur la mesure de l'évolution des inégalités scolaires. Le dernier article paru sur ce sujet est celui de Louis-André Vallet (2007).
4. L'écriture résumée d'un modèle logit multinomial peut être la suivante :

$$\ln\left(\frac{\mathrm{P_r}(Y=m)}{\mathrm{P_r}(Y=M)}\right)=\sum_{r=1}^{R}\beta_{mr}X_r$$

   avec Y modalités de la variable à expliquer, variant de 1 à M (la modalité M est mise par convention en référence), X_r variables explicatives introduites dans le modèle, variant de 1 à R et $X_1=1$ (constantes du modèle), et β correspondant aux M*r coefficients introduits dans le modèle.
5. Pour un exemple d'utilisation, voir Deauvieau et Dumoulin (2010).
6. A ce titre, une modélisation logit n'est qu'un cas particulier de la modélisation log linéaire. Cette façon de présenter les choses est très courante dans le monde anglo-saxon mais beaucoup moins en France; voir par exemple l'ouvrage de Demaris (1990). Il existe aux Etats-Unis depuis les années 1940 un courant de recherche sur la modélisation des variables qualitatives qui n'a pas vraiment vu le jour en France, en tout cas dans les univers des sociologues (on trouvera des éléments d'appréciation sur ce point dans Vallet, 2010). L'une des explications possibles à cet état de fait est peut-être qu'historiquement le courant de l'analyse des donnés en France a probablement joué le rôle du courant de la modélisation des tableaux croisés au Etats-Unis et conjointement que la régression logistique est introduite en France par l'INSEE et donc d'abord envisagée comme une méthode économétrique particulière issue de l'univers des données numériques.

References

Deauvieau J (2010) Présenter les résultats d'une modélisation logit sous forme de probabilités. *Bulletin de Méthodologie Sociologique,* 105 : 5-23.

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

Demaris A (1990) *Logit Modeling. Pratical Applications*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-086. Newbury Park, CA : Sage.

Gujarati D (2004) *Econométrie*, Paris : De Boeck.

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

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*, 111: 76-79.

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.

Vallet LA (2010) Quelques aspects de la statistique en sociologie, 1950–2010. *Mathématiques et Sciences humaines*, 191 : 65-80.