Trideux Software Integrates Jérôme Deauvieau's "How to Translate a Logit Model into Probabilities", *Bulletin de Méthodologie Sociologique* 105, 2010 p.53-60 (version française)

Evolution du logiciel Trideux pour prendre en compte les résultats de l'article de Jérôme Deauvieau 'Comment traduire sous forme de probabilités les résultats d'une modélisation logit?'

Philippe Cibois

Laboratoire Printemps CNRS / UVSQ

phcibois@wanadoo.fr

Résumé

Jérôme Deauvieau dans son article "Comment traduire sous forme de probabilités les résultats d'une modélisation logit ?" montre que la présentation sous forme de probabilités présente de défaut de varier selon le choix de la catégorie de référence. Il présente deux méthodes qui permettent d'éviter ces difficultés, l'ajustement selon l'écart expérimental et l'ajustement selon l'écart pur. Ces deux méthodes sont maintenant implémentées dans le logiciel Trideux.

Dans son article du présent numéro du BMS "Comment traduire sous forme de probabilités les résultats d'une modélisation logit ?", Jérôme Deauvieau souligne le fait que les sociologues (en France en particulier) utilisent souvent dans leurs données et donc dans leurs régressions logistiques, des variables discrètes et non des variables continues. Pour lui la raison en est que les sociologues " utilisent très souvent des variables explicatives catégorielles (sexe, PCS…), et lorsqu'ils sont en présence de variables numériques (salaire, âge…), l'usage courant veut qu'elles soient mises en catégories."

Cette manière de voir a des conséquences dans la manière d'exposer les résultats de régressions logistiques faites sur ces données discrètes. En effet puisque la logique de présentation habituelle des variables discrètes est le tableau croisé où sont repérés les écarts de pourcentages, on reste bien dans cette logique de présentation en donnant les résultats de régressions logistiques sous forme de probabilités ou de pourcentages. De plus, comme ces résultats sont plus simples à lire que les coefficients d'un modèle logit ou que des Odds ratios, il est possible ainsi de communiquer les résultats obtenus à un plus vaste public.

Les difficultés de la présentation en probabilité

Cette présentation en probabilité ou en pourcentage pose cependant des problèmes car les résultats différent selon le choix fait des modalités qui servent de référence et Deauvieau propose deux solutions pour résoudre cette difficulté.

Dans ce qui suit, je montrerai comment ces méthodes sont utilisables désormais dans le logiciel Trideux : dans un premier temps, et pour montrer les difficultés de la représentation habituelle, je présenterai un exemple concernant des données concernant le choix de l'étude du latin, option facultative qui peut être prise en deuxième année du cursus secondaire en France. Les données utilisées sont celle du Panel 1995 du Ministère de l'Education nationale.

L'initiation au latin a lieu en classe de 5e et elle est suivie par 27,8% des élèves. Ce sont des bons élèves dès le primaire : à leur arrivée en 6e, leur niveau a été apprécié dans quatre domaines (lecture, français écrit et oral, mathématiques) et une note leur a été donnée allant de 0 à 10. Prenons le fait d'avoir de 8 à 10 comme mention d'excellence : pour la population de 5e en général 17% ont une mention d'excellence dans toutes les matières alors que 34% des latinistes sont dans ce cas.

Les latinistes se différencient des autres par un certain nombre d'aspects :

- choix de la première langue : ils sont deux fois plus nombreux que les non-latinistes (18,0% contre 9,8%) à avoir choisi l'allemand en première langue en 6e. On sait qu'il s'agit souvent d'une stratégie pour être dans une "bonne classe" (sauf dans les régions proches de l'Allemagne),
- leurs parents sont plus souvent cadres ou professions intellectuelles supérieures (28,3% contre 10,4%), professions intermédiaires (23,1% contre 16,1%), mais moins employés (14,8% contre 18,7%) ou ouvriers (19,9% contre 38,4%). De même leurs parents sont plus nombreux à avoir fait des études supérieures (père : 28,2% contre 10,0% ; mère : 30,7% contre 10,3%),
- cours de musique : ils sont également deux fois plus nombreux (19,5% contre 8,7%) à prendre régulièrement des cours de musique en dehors de l'établissement scolaire.

Pour savoir ce qui a eu le plus d'influence sur le choix du latin parmi les variables indiquées précédemment, il est possible de faire des comparaisons "toutes choses égales par ailleurs" (régression logistique). Les résultats issus de Trideux sont les suivants (tous les effets sont significatifs au seuil de 1%) :

Paramètres de la régression **en pourcentages**
Régression logistique
Modalité à expliquer : Choix du latin en 5e
Situation de référence Pas de mention d'excellence
Anglais en 6e
Père sans dipl. sup
Mère sans dipl. sup
Pas de cours de musique

	Paramètres	ChancesRef	%
	-1.5980	0.2023	16.8
Effets marginaux		**Odds-ratio**	
Mentions excellence en 6e	1.2519	3.50	24.6
Allemand en 6e	0.4464	1.56	7.2
Père diplôme supérieur	0.6261	1.87	10.6
Mère diplôme supérieur	0.9152	2.50	16.7
Cours de musiques	0.5365	1.71	8.9

Comme on a pris comme situation de référence tout ce que les tris croisés ont montré comme n'allant pas dans le sens du choix de l'option latin, l'estimation en pourcentage de la situation de référence (16,8%) est loin de la situation observée (27,8%). On voit que l'effet le plus important est lié aux mentions d'excellence qui augmentent la proportion de latiniste de 25%, puis le diplôme supérieur de la mère, celui du père, le fait d'avoir pris l'allemand en 6e ou des cours de musique. Il est bien certain que cette distinction d'effets est artificielle car l'ensemble de ces caractéristiques forme une logique éducative des familles d'origine sociale supérieure. Cependant la régression logistique met en avant le résultat de cette logique : l'excellence scolaire est primordiale, même si d'autres éléments comme la

culture musicale font partie des objectifs de formation. Le fait que la musique ait un "effet" sur le choix du latin manifeste que ce qui est visé dans l'ensemble du comportement éducatif est la recherche de l'excellence sociale.

Dans la logique de la critique de Deauvieau de ce résultat, changeons de situation de référence en modifiant simplement la première et en prenant le fait d'avoir des mentions d'excellence en 6e. Ceci correspond à une situation de référence ou le choix de l'option latin est plus fréquent et ceci va donc faire monter le pourcentage du choix de la situation de référence.

Paramètres de la régression **en pourcentages**
Régression logistique
Modalité à expliquer : Choix du latin en 5e
Situation de référence Mentions excellence en 6e
Anglais en 6e
Père sans dipl. sup
Mère sans dipl. sup
Pas de cours de musique

	Paramètres	ChancesRef	%
	-0.3461	0.7074	41.4
Effets marginaux		**Odds-ratio**	
Absence d'excellence	-1.2519	0.29	-24.6
Allemand en 6e	0.4465	1.56	11.1
Père diplôme supérieur	0.6262	1.87	15.5
Mère diplôme supérieur	0.9152	2.50	22.4
Cours de musiques	0.5366	1.71	13.3

La situation de référence est comme prévu au-dessus de la situation moyenne, tous les Odds ratios sont identiques sauf celui de la première question dont on a maintenant l'inverse : 0,29 = 1 / 3,5. Pour cette question, l'effet en pourcentage est simplement changé de signe mais tous les autres effets sont différents et c'est cette prise de conscience de la non stabilité des effets en pourcentage qui conduit Deauvieau à proposer deux méthodes qui permettent de donner des résultats sous forme de probabilités ou de pourcentages quelque soit la situation de référence prise.

Première solution : l'écart expérimental

Reprenons les termes de Deauvieau : "on calcule pour chaque individu de l'échantillon sa probabilité individuelle de connaître l'évènement modélisé. A partir de ces probabilités individuelles, on calcule des probabilités « théoriques » en assumant la posture « expérimentale » inhérente aux méthodes de régression multiple. En effet la base épistémologique de la régression multiple consiste à découper les individus selon leurs différentes caractéristiques sociales, et à mesurer l'effet d'une des caractéristiques indépendamment des autres. Cela revient en fait à imiter le raisonnement expérimental qu'on trouve par exemple en biologie : je cherche à mesurer l'effet d'une caractéristique sur un phénomène donné, et pour cela je fais une expérience consistant à donner cette caractéristique à un groupe, à omettre cette caractéristique pour un second groupe, et à regarder ce qui se passe en comparant les deux groupes. (…)"

"Le jeu de coefficients estimés par le modèle permet précisément de se mettre par le calcul dans une posture expérimentale. Prenons par exemple la variable sexe

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

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

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

$$P\ (Y=cadre) = \frac{1}{1+\exp^{-(-1,96+0,75*diplome-0,29*age2-0,63*age3)}}$$

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

$$P\ (Y=cadre) = \frac{1}{1+\exp^{-(-1,96+0,75*diplome+0,59-0,29*age2-0,63*age3)}}$$

"L'inconvénient principal de ce mode de présentation est que le résultat obtenu n'est pas forcément égal au contraste logistique tel qu'il est indiqué par le coefficient logit du modèle", ce qui n'est pas le cas de la méthode présentée maintenant.

Deuxième solution : l'écart "pur"

Il s'agit de la solution proposée par Léridon et Toulemon (1997 : 251). Le principe consiste à trouver les probabilités associées aux modalités d'une variable explicative qui satisfassent aux deux conditions suivantes :

1/ l'écart entre les probabilités respecte le contraste logistique entre les modalités tel qu'il est défini par le coefficient du modèle.

2/ la moyenne pondérée des probabilités liées aux modalités de la variable explicative est égale à la probabilité moyenne de la variable à expliquer sur l'ensemble de l'échantillon.

Soit *p* la proportion moyenne observée de la variable à expliquer.

Soit une variable explicative possédant *n* modalités numérotées de *1* à *n*. Chaque modalité est d'effectif n_i, de proportion p_i, *i* variant de *1* à *n*.

Soit $alpha_i$ les paramètres de la régression logistique pour cette variable explicative, $alpha_n$ correspondant à la modalité de référence ($alpha_n$ = 0, exp($alpha_n$)=1)

soit $Pajust_i$ les proportions ajustées recherchées (celle d'indice *n* étant celle de la modalité de référence). La condition 1 impose que :

$Pajust_i$ /(1- $Pajust_i$) / $Pajust_n$/ (1- $Pajust_n$) = exp($alpha_i$) d'où l'on tire :

si l'on pose A= exp($alpha_i$) x $Pajust_n$ / (1- $Pajust_n$) alors :

$Pajust_i$ = A / (1 + A)

La deuxième condition est que la somme des n_i x $Pajust_i$ soit égale à p, la proportion moyenne observée.

L'algorithme (en encadré) consiste à choisir $Pajust_n$ arbitrairement puis à modifier cette valeur par ajout ou soustraction des valeurs successives de 2^{-k} (k variant de 1 à une valeur suffisante pour la précision recherchée, 30 par exemple) selon que le résultat de la deuxième condition est inférieur ou supérieur à la proportion moyenne observée.

Pajustn désigne $Pajust_n$, Effectif(i) désigne n_i, alpha(i) désigne $alpha_i$ VarA est l'intermédiaire de calcul A, SommeNP désigne la somme des n_i x $Pajust_i$ dont PropSommeNP est la proportion par rapport au total de la modalité à expliquer totalVarY, PropMoyenneObs est p, la moyenne générale de la modalité à expliquer. La valeur de la variable ajustée pour la $i^{ème}$ modalité se trouve à la fin dans VarAjust(i)

```
Pajustn = 0
For K = 1 To 30
    SommeNP = 0
    Pajustn = Pajustn + 2 ^ -K
    For i = 1 To nmax
        VarA = exp(alpha(i)) * Pajustn / (1 - Pajustn)
        VarAjust(i) = VarA / (1 + varA)
        SommeNP = SommeNP + VarAjust(i) * Effectif(i)
    Next i
    PropSommeNP = SommeNP / totalVarY
    If PropMoyenneObs - PropSommeNP < 0 Then
         Pajustn = Pajustn - 2 ^ -K
    End If
Next K
```

Résultats des deux méthodes

Reprenons les données présentées plus haut. Cette fois, quelque soit la situation prise comme référence, les proportions ajustées sont identiques et cela est vrai pour les deux méthodes.

```
Sorties en probabilité selon méthode de l'effet expérimental (Deauvieau,
BMS)
Modalité à expliquer : Choix du latin en 5e, moyenne générale 0.2777

MoyExp : Moyenne expérimentale
Ecart1 : Ecart de la moyenne expérimentale à la moyenne générale en %
Ecart2 : Ecart de la moyenne observée à la moyenne générale en %
                         MoyExp    Ecart1   Ecart2
Mentions excellence en 6e 0.4913    21.4    28.6
Absence  d'excellence     0.2335    -4.4    -5.8
Allemand en 6e            0.3496     7.2    16.3
Anglais en 6e             0.2679    -1.0    -2.1
Père diplôme supérieur    0.3787    10.1    28.1
Père sans dipl.sup.       0.2593    -1.8    -4.6
Mère diplôme supérieur    0.4309    15.3    29.5
```

```
Mère sans dipl.sup.       0.2487    -2.9    -5.2
Cours de musiques         0.3660     8.8    21.1
Pas de cours musique      0.2660    -1.2    -2.6

Sorties en probabilité selon méthode de l'effet pur (Leridon/Toulemon 1997
p.252)
Modalité à expliquer : Choix du latin en 5e, moyenne générale 0.2777

MoyAjust : Moyenne ajustée selon l'effet pur
Ecart1 : Ecart de la moyenne ajustée à la moyenne générale en %
Ecart2 : Ecart de la moyenne observée à la moyenne générale en %
                         MoyAjust Ecart1   Ecart2
Mentions excellence en 6e 0.5115   23.4    28.6
Absence  d'excellence     0.2304   -4.7    -5.8
Allemand en 6e            0.3625    8.5    16.3
Anglais en 6e             0.2668   -1.1    -2.1
Père diplôme supérieur    0.3948   11.7    28.1
Père sans dipl.sup.       0.2586   -1.9    -4.6
Mère diplôme supérieur    0.4508   17.3    29.5
Mère sans dipl.sup.       0.2474   -3.0    -5.2
Cours de musiques         0.3812   10.4    21.1
Pas de cours musique      0.2648   -1.3    -2.6
```

Pour les deux méthodes et pour toutes les modalités explicatives on donne la moyenne expérimentale ou la moyenne ajustée en proportion et, dans la colonne "Ecart1", en pourcentage l'écart à la moyenne générale, ici la proportion du choix du latin en 5^{e} soit 27,7% de la population. On constate immédiatement que les deux méthodes donnent des résultats proches.

La colonne marquée "Ecart2", qui est identique pour les deux méthodes donne un résultat de simple tri croisé entre le pourcentage en ligne de la modalité et le pourcentage toutes lignes confondues, c'est-à-dire toujours la même moyenne générale. Quand pour toutes les modalités d'une question, les écarts donnés par la régression logistique et les écarts donnés par les tris croisés sont très proches, cela signifie que la procédure "toutes choses égales par ailleurs" n'apporte pas plus que les tris croisés. Ce n'est pas le cas ici car les différences peuvent être importantes.

A titre de comparaison entre d'une part les deux analyses précédentes où les situations de références étaient différentes et d'autre part la méthode de la moyenne ajustée selon l'effet pur, classons les modalités par ordre d'importance d'effet.

1) Première expérience (*cf.* plus haut)

```
Mentions excellence en 6e          24.6
Mère diplôme supérieur             16.7
Père diplôme supérieur             10.6
Cours de musiques                   8.9
Allemand en 6e                      7.2
```

2) Deuxième expérience *(idem.* Ici on classe par valeur absolue)

```
Absence  d'excellence             -24.6
Mère diplôme supérieur             22.4
Père diplôme supérieur             15.5
Cours de musiques                  13.3
Allemand en 6e                     11.1
```

3) Méthode selon l'effet pur : on classe en prenant les différences entre les modalités.

Mentions excellence en 6e	**0.5115**	
Absence d'excellence	**0.2304**	
Différence		**28.1**
Mère diplôme supérieur	**0.4508**	
Mère sans dipl.sup.	**0.2474**	
Différence		**20.3**
Père diplôme supérieur	**0.3948**	
Père sans dipl.sup.	**0.2586**	
Différence		**13.6**
Cours de musiques	**0.3812**	
Pas de cours musique	**0.2648**	
Différence		**11.7**
Allemand en 6e	**0.3625**	
Anglais en 6e	**0.2668**	
Différence		**9.6**

On retrouve le même ordre et des ordres de grandeur proches mais on dispose maintenant d'une méthode qui permet ne plus se poser la question du choix de la référence, ce qui est particulièrement rassurant.

Bibliographie

Deauvieau, Jérôme, (2010), Comment traduire sous forme de probabilités les résultats d'une modélisation logit ?, *BMS* 105, Janvier 2010.

Leridon, Henri et Toulemon, Laurent, (1997), *Démographie, approche statistique et dynamique des populations*, Paris, Economica.