Bulletin de Méthodologie Sociologique N°64 Octobre 1999 - p.5-24

Modèle linéaire contre modèle logistique en régression sur données qualitatives

Philippe Cibois
Département de sociologie,
Université de Versailles - St-Quentin
phcibois@wanadoo.fr

Abstract. **The Linear Model Versus The Logistic Model in Regression on Qualitative Data.** Regression on qualitative data is usually done by using a logistic model. By examining data where "everything is otherwise equal" one can show that the linear model is quite compatible with this type of data. Results of an in-deep analysis of cross-tabulation data (called *tabular analysis*), and of regression using linear and logistic models, are quite similar (in terms of percentage effects). The theoretical question concerning the possibility of a linear model remains to be examined : it seems that the specific situation of a reference category introduces constraints which make the linear model possible. **Qualitative Data, Tabular Analysis, Logistic Regression, Linear Regression.**

Résumé. La régression sur données qualitatives est habituellement traitée en utilisant un modèle logistique. En examinant des données "toutes choses égales par ailleurs", on montre que le modèle linéaire est tout à fait cohérent avec ce type de données. En comparant les résultats d'une analyse en profondeur des tableaux croisés (appelée *analyse tabulaire*) avec la régression linéaire et la régression logistique, on constate que ces diverses méthodes donnent des résultats très proches (en termes d'effets en pourcentages). La question théorique de la possibilité du modèle linéaire reste à approfondir : le cas particulier des écarts à une situation de référence introduit des contraintes qui semblent rendre possible l'utilisation du modèle linéaire. **Données qualitatives, Analyse tabulaire, Régression logistique, Régression linéaire**

La régression sur données qualitatives a d'abord été pratiquée dans des milieux scientifiques qui traitaient des données biologiques, variables quantitatives et qualitatives mélangées : démographie, écologie, épidémiologie. Préalablement à la régression, le modèle logistique mis au point par Verhulst l'a été dans le cadre de la démographie au 19^{e} siècle et celui qui a inventé le terme *logit* (Berkson 1944) faisait référence à la biologie dans le titre de son article. Si l'on examine plus précisément la régression logistique en épidémiologie (Bouyer 1991 : 80), on constate qu'elle sert à modéliser la relation entre une variable qualitative en présence/absence et des variables qui peuvent être quantitatives ou qualitatives, selon un modèle mis au point par Cox (1972). Le modèle a ensuite été repris par les économistes (Gouriéroux 1989 : 3) où les variables à expliquer sont qualitatives mais où les variables explicatives peuvent encore être un mélange de qualitatifs et de quantitatifs[1]. Par

[1] par exemple (p.27) réussite ou non à un examen expliquée par des variables quantitatives (taille de la commune, ressources des parents, âge, note à un test, moyenne au bac) et des variables qualitatives (type de bac, vient du public ou non, première année ou non de supérieur).

contre dans un ouvrage collectif plus récent de l'INSEE (Lollivier 1996), ne sont plus envisagés que des modèles où toutes les variables sont qualitatives, la variable à expliquer évidemment, mais aussi les autres.

C'est de ce seul cas que nous traiterons dans la suite : nous souhaitons trouver un modèle pour des données où toutes les variables soient nominales. D'un point de vue formel, nous sommes dans le cas d'un tableau de contingence, si nous considérons les tableaux croisés obtenus à partir des données ; mais nous sommes dans le cas d'une base de données de variables nominales si nous considérons les données de base qui permettent de construire ces tableaux croisés où à chaque individu et pour chaque variable correspond soit le numéro d'une modalité (codage ordinaire) soit un codage en présence/absence (codage disjonctif complet).

Dans ce qui suit, nous allons d'abord, en suivant l'adage que *le modèle doit suivre les données et non l'inverse*[2], examiner sur des exemples la procédure clé de la régression sur variables qualitatives, c'est à dire la mise en relief d'effets "toutes choses égales par ailleurs". Nous partons en effet de l'hypothèse que dans un tableau croisé où plusieurs variables sont croisées, l'action "toutes choses égales par ailleurs" d'une variable sur une autre doit se repérer pour tous les cas où toutes les autres modalités sont identiques. Le travail à faire devient de ce fait une procédure de neutralisation d'une ou plusieurs variables comme on en fait en analyse multivariée.

Expliquer la lecture

A titre d'exemple, on utilise les données de l'enquête sur *les pratiques culturelles des français* de 1989[3]. La variable que l'on va chercher à expliquer sera la lecture, repérée par les réponses suivantes à la question :

Au total, diriez-vous que vous êtes plutôt quelqu'un qui lit 1) beaucoup de livres, 2) moyennement, 3) peu, 4) pas.

On cherchera à rendre compte de cette auto-estimation par le sexe et le diplôme.

Pour simplifier le problème on dichotomise chaque variable : les niveaux "beaucoup" et "moyen" de lecture sont rassemblés en FORT, les niveaux "peu" ou "pas" en FAIBL. Les diplômes inférieurs au bac sont par convention notés NDIP, le bac, les étudiants et élèves en cours d'études et les diplômes supérieurs sont notés par convention BACS.

La distribution de chaque question est la suivante :

```
Question SEXE          Question DIPLOME     Question LECTURE
Tot. MASC FEMI         Tot. NDIP BACS       Tot. FORT FAIBL
4997 2404 2593         4997 3402 1595       4997 2512 2485
 100 48.1 51.9          100 68.1 31.9        100 50.3 49.7
```

[2] "Bien mieux qu'à des modèles conjecturaux, c'est à l'observation qu'on doit demander quel est l'ordre de la réalité : le mérite du calculateur étant de découvrir sans parti pris, sans *a priori*, quels courants de lois traversent l'océan des faits. " J.-P. Benzécri (1976 I : V) mis sous forme d'adage par H. Rouanet (1993 : VI)

[3] *Nouvelle enquête sur les pratiques culturelles des français en 1989*, La Documentation française, 1990.

Pour pouvoir juger "toutes choses égales par ailleurs", il faut trier en profondeur de façon à obtenir des catégories équivalentes pour toutes les variables indépendantes. On regroupe donc pour chaque niveau de chaque variable tous les autres niveaux des autres variables. On constitue donc des lignes "MASC - NDIP", "MASC-BACS", "FEMI-NDIP", "FEMI-BACS", qui représentent la totalité des cas de figures de croisement lexicographique des modalités explicatives entres elles. Ce sont tous ces cas que l'on croise avec la variable à expliquer. On a les résultats suivants en effectifs et en pourcentages :

		FORT	FAIBL	tot	FORT	FAIBL	tot
1	MASC-NDIP	619	1009	1628	38.0	62.0	100
2	*MASC-BACS*	*478*	*298*	*776*	*61.6*	*38.4*	*100*
3	FEMI-NDIP	794	980	1774	44.8	55.2	100
4	FEMI-BACS	621	198	819	75.8	24.2	100
			Total	4997			

Prenons comme référence, comme point de départ des comparaisons, le fait d'être de sexe masculin et d'avoir un diplôme égal ou supérieur au bac. Cette situation de référence correspond à la deuxième ligne du tableau et nous voyons que dans ce cas le taux de fort lecteur est de 61,6%.

Examinons l'effet de non-diplôme, effet qui peut être repéré en neutralisant l'effet de sexe : ceci est obtenu en prenant d'abord dans la population de sexe masculin les deux sous-populations suivantes qui ne s'opposent que par le diplôme :

		FORT	FAIBL	tot	FORT	FAIBL	tot
1	MASC-NDIP	619	1009	1628	38.0	62.0	100
2	MASC-*BACS*	478	298	776	61.6	38.4	100

La différence de pourcentage de fort lecteur, si nous prenons la situation de diplômé comme référence, est de 38,0 - 61,6 = -23,6 : c'est l'effet marginal du non-diplôme dans la population masculine.

Pour la population de sexe féminin, le même effet de non-diplôme est de :

		FORT	FAIBL	tot	FORT	FAIBL	tot
3	FEMI-NDIP	794	980	1774	44.8	55.2	100
4	FEMI-*BACS*	621	198	819	75.8	24.2	100

L'effet marginal est de 44,8 - 75,8 = -31,1[4]. Les deux effets sont négatifs et importants : l'absence de diplôme ou un faible diplôme n'encourage pas à la lecture, quelque soit le sexe. Si l'on veut simplifier le problème on considérera les deux effets comme proches et on pourra prendre leur moyenne : (-23,6 -31,1)/2 = -27,3.

On peut même raffiner en considérant que chaque différence est apportée par des effectifs différents, la population masculine de 1628+776 soit 2404 hommes sur 4997 répondants au total (donc une proportion de 2404/4997 = 0,481). On a de même 1774 + 819 soit 2593 femmes, 0,519 en proportion.

On pondère donc les deux effets pour avoir un effet moyen de :

(-23,6 x 0,481) + (-31,1 x 0,519) = -27,5

[4] Ici et dans la suite, les calculs sont faits avec la précision maximum mais tous les affichages sont faits avec une précision de un chiffre après la virgule.

Examinons ensuite l'effet de sexe féminin, effet qui peut être repéré en neutralisant l'effet de diplôme : dans la population des sans diplôme on a les deux sous-populations suivantes qui ne s'opposent que par le sexe :

	FORT	FAIBL	tot	FORT	FAIBL	tot
1 *MASC*-NDIP	619	1009	1628	38.0	62.0	100
3 FEMI-NDIP	794	980	1774	44.8	55.2	100

La différence de pourcentage de fort lecteur, si nous prenons le sexe masculin comme référence, correspond à un effet marginal de 44,8 - 38,0 = 6,7

Dans la population des diplômés, le même effet de sexe est de :

	FORT	FAIBL	tot	FORT	FAIBL	tot
2 *MASC*-BACS	478	298	776	61.6	38.4	100
4 FEMI-BACS	621	198	819	75.8	24.2	100

75,8 - 61,6 = 14,2. On voit que l'effet du sexe féminin est dans les deux cas positif (les femmes lisent plus que les hommes), plus faible dans les cas des non-diplômés. Toujours avec la même simplification, la moyenne pondérée des effets est cette fois de 9,1.

Résumons nous : la situation de référence (MASC, BACS) correspond à un pourcentage de fort lecteur de 61,6. L'effet de sexe féminin est de +9,1 ; l'effet de faible diplôme est de -27,5 En cumulant éventuellement les effets (hypothèse simplificatrice), on peut ainsi comparer les situations observées et les situations estimées de pourcentage de forts lecteurs :

Situation	Effets	Observation	Estimation (ref + effet)	Erreur (Obs-est)
2 MASC BACS	référence	61,6	61,6	0,0
4 FEMI BACS	effet FEMI	75,8	61,6 + 9,1 = 70,7	5,1
1 MASC NDIP	effet NDIP	38,0	61,6 −27,5 = 34,1	3,9
3 FEMI NDIP	FEMI & NDIP	44,8	61,6 + 9,1 −27,5 = 43,3	1,5

On voit que, malgré les simplifications apportées, les estimations sont assez proches de la réalité.

Les simplifications qui ont été faites ont consisté à prendre l'effet moyen à la place des sous-effets qui étaient de même signe et de même ordre de grandeur. Dans l'esprit de la régression on voit que si on prend une référence et qu'on lui ajoute des effets, on a un modèle purement additif qui n'est pas trop éloigné des données.

Comme ces données sont extrêmement simples, il est semble prudent de voir si sur des données plus complexes, le même modèle additif, pour le moment empirique, semble réaliste. Nous introduisons, en plus du sexe et du diplôme, l'âge et la catégorie socioprofessionnelle (CSP).

Expliquer la lecture par le sexe , l'âge, le diplôme et la CSP

On recode l'âge en trois positions : *jeunes* de moins de 25 ans, âge *médian* pour 25-49, *âgés* pour 50 et plus. Le diplôme est recodé comme précédemment en faiblement diplômé ou non. Pour la CSP on met dans la classe supérieure la catégorie cadres et professions intellectuelles supérieures ainsi que les professions intermédiaires, le reste étant mis dans la classe inférieure. Ici la lecture est

catégorisée en prenant comme niveau fort, ceux qui lisent beaucoup (ce qui réduit leur nombre par rapport à l'exemple précédent). On a les tris à plat suivants :

```
Question SEX          Question AGE                Question DIP
Tot. MASC FEMI        Tot. JEUN MEDI AGEE         Tot. NDIP BACS
4997 2404 2593        4997 1000 2273 1724         4997 3402 1595
 100 48.1 51.9         100 20.0 45.5 34.5          100 68.1 31.9

Question CSP          Question LEC
Tot. CSUP CINF        Tot.  BCP AUTR
4997 1172 3825        4997  834 4163
 100 23.5 76.5         100 16.7 83.3
```

Les données de base sont données ci-dessous : pour chacun des croisements possibles de chaque modalité explicative, on donne l'effectif observé et le pourcentage de la variable à expliquer (pourcentage de très forts lecteurs). On notera que l'on nomme *données de base* le tableau ci-dessous parce qu'il permet de calculer tous les sous-effets. D'un point de vue formel c'est un simple tri croisé entre la variable à expliquer et toutes les situations de variables explicatives possibles (et toutes les situations théoriquement possible peuvent ne pas être attestées : cela diminue la fiabilité des résultats mais ne rend pas les calculs impossibles).

Par exemple la première ligne indique que les homme jeunes non-diplômés de classe supérieure sont 33 et que parmi eux seulement 12,1% se considèrent très forts lecteurs. Le "R" indique les modalités qui servent de référence : comme le choix est arbitraire et sans conséquences, on a pris le plus fort effectif de la ligne 22 pour que la situation de référence soit statistiquement bien déterminée.[5]

n°	Situation				Effectif	% de forts lecteurs
1	MASC	JEUN	NDIP R	CSUP	33	12.1
2	MASC	JEUN	NDIP R	CINF R	190	8.4
3	MASC	JEUN	BACS	CSUP	92	14.1
4	MASC	JEUN	BACS	CINF R	172	9.3
5	MASC	MEDI	NDIP R	CSUP	124	14.5
6	MASC	MEDI	NDIP R	CINF R	581	8.8
7	MASC	MEDI	BACS	CSUP	222	29.7
8	MASC	MEDI	BACS	CINF R	121	19.8
9	MASC	AGEE R	NDIP R	CSUP	36	19.4
10	MASC	AGEE R	NDIP R	CINF R	664	9.6
11	MASC	AGEE R	BACS	CSUP	45	40.0
12	MASC	AGEE R	BACS	CINF R	124	26.6
13	FEMI R	JEUN	NDIP R	CSUP	18	11.1
14	FEMI R	JEUN	NDIP R	CINF R	207	12.1
15	FEMI R	JEUN	BACS	CSUP	94	38.3
16	FEMI R	JEUN	BACS	CINF R	194	22.7
17	FEMI R	MEDI	NDIP R	CSUP	172	16.3
18	FEMI R	MEDI	NDIP R	CINF R	622	12.2
19	FEMI R	MEDI	BACS	CSUP	257	44.4
20	FEMI R	MEDI	BACS	CINF R	174	27.6
21	FEMI R	AGEE R	NDIP R	CSUP	51	19.6

[5] Avec un effectif de 704, l'intervalle de confiance sur le pourcentage de 12,4 est, à un écart-type, de ± 1,2%

```
22 FEMI R  AGEE R  NDIP R  CINF R    704 12.4
23 FEMI R  AGEE R  BACS    CSUP       28 35.7
24 FEMI R  AGEE R  BACS    CINF R     72 33.3
```

Il y a 24 cas à envisager : 2 modalités de sexe multipliées par 3 d'âge, par 2 de diplôme et 2 de CSP et ici tous ces cas sont attestés.

La situation de référence est la suivante :

```
22 FEMI R  AGEE R  NDIP R  CINF R    704 12.4
```

On calcule maintenant l'effet de chaque modalité qui n'est pas de référence. Par exemple pour la modalité de sexe masculin, le premier effet est calculé en comparant la ligne 1 des données et la ligne 13, identiques en tout sauf pour le sexe. L'effet est calculé par différence entre les deux taux de lecteurs :

```
 1 MASC     JEUN    NDIP R  CSUP       33 12.1
13 FEMI R   JEUN    NDIP R  CSUP       18 11.1
```

La différence est de 1,0 et concerne un effectif de 33 + 18 = 51 jeunes non-diplômés de classe supérieure, qu'ils soient de sexe masculin (33) ou féminin (18).

Il y a 11 autres effets (2 - 14, 3 - 15, 4 - 16, etc.), qui correspondent à des situations toutes identiques, sauf pour le sexe, dont la moyenne pondérée par ce que représente chaque effectif par rapport au total est de -5,9[6]. Les autres effets à deux modalités se calculent de manière analogue. On donne pour chaque ligne les caractéristiques de la sous-population sur laquelle est calculée l'effet, la valeur de l'effet, l'effectif de la sous-population. Enfin on calcule la moyenne pondérée et l'écart-type pondéré pour se rendre compte de la dispersion.

```
Effet MASC  12  effets
 JEUN NDIP CSUP   1.0   51
 JEUN NDIP CINF  -3.7  397
 JEUN BACS CSUP -24.2  186
 JEUN BACS CINF -13.4  366
 MEDI NDIP CSUP  -1.8  296
 MEDI NDIP CINF  -3.4 1203
 MEDI BACS CSUP -14.6  479
 MEDI BACS CINF  -7.8  295
 AGEE NDIP CSUP  -0.2   87
 AGEE NDIP CINF  -2.7 1368
 AGEE BACS CSUP   4.3   73
 AGEE BACS CINF  -6.7  196
               Total 4997

moyenne pondérée des effets = -5.9
Ecart-type pondéré          =  5.7
```

Comme la division de la question âge est à trois modalités, l'effet jeune et l'effet âge médian se calculent de façon semblable, mais il n'y a que 8 effets et les pondérations sont calculées sur une base différente[7].

---

[6] calcul de la moyenne pondérée = 1,0 x (51 / 4997) - 3.7 x (397 / 4997) + etc.

[7] Il n'y a que 8 effets car les situations semblables hors effet d'âge sont de 2 modalités de sexe multipliée par 2 modalités de diplômes et 2 modalités de CSP. Le dénominateur de la pondération est

```
Effet JEUN  8  effets
 MASC NDIP CSUP  -7.3   69
 MASC NDIP CINF  -1.2  854
 MASC BACS CSUP -25.9  137
 MASC BACS CINF -17.3  296
 FEMI NDIP CSUP  -8.5   69
 FEMI NDIP CINF  -0.3  911
 FEMI BACS CSUP   2.6  122
 FEMI BACS CINF -10.7  266
               Total 2724
moyenne pondérée des effets = -5.0
Ecart-type pondéré          =  7.5

Effet MEDI  8  effets
 MASC NDIP CSUP  -4.9  160
 MASC NDIP CINF  -0.9 1245
 MASC BACS CSUP -10.3  267
 MASC BACS CINF  -6.8  245
 FEMI NDIP CSUP  -3.3  223
 FEMI NDIP CINF  -0.1 1326
 FEMI BACS CSUP   8.6  285
 FEMI BACS CINF  -5.7  246
               Total 3997
moyenne pondérée des effets = -1.5
Ecart-type pondéré          =  4.1


Effet BACS  12  effets
 MASC JEUN CSUP   2.0  125
 MASC JEUN CINF   0.9  362
 MASC MEDI CSUP  15.2  346
 MASC MEDI CINF  11.1  702
 MASC AGEE CSUP  20.6   81
 MASC AGEE CINF  17.0  788
 FEMI JEUN CSUP  27.2  112
 FEMI JEUN CINF  10.6  401
 FEMI MEDI CSUP  28.1  429
 FEMI MEDI CINF  15.4  796
 FEMI AGEE CSUP  16.1   79
 FEMI AGEE CINF  21.0  776
               Total 4997
moyenne pondérée des effets = 15.6
Ecart-type pondéré          =  6.8

Effet CSUP  12  effets
 MASC JEUN NDIP   3.7  223
```

---

toujours le complément au total de la somme de la modalité de référence et de la modalité dont on étudie l'effet. Dans le cas dichotomique, c'est donc l'effectif total, ce ne l'est plus dans les autres cas.

```
 MASC JEUN BACS    4.8  264
 MASC MEDI NDIP    5.7  705
 MASC MEDI BACS    9.9  343
 MASC AGEE NDIP    9.8  700
 MASC AGEE BACS   13.4  169
 FEMI JEUN NDIP   -1.0  225
 FEMI JEUN BACS   15.6  288
 FEMI MEDI NDIP    4.1  794
 FEMI MEDI BACS   16.8  431
 FEMI AGEE NDIP    7.2  755
 FEMI AGEE BACS    2.4  100
                Total 4997
moyenne pondérée des effets =  7.8
Ecart-type pondéré          =  4.5
```

En résumé on a :

```
Situation de référence : FEMI AGEE NDIP CINF =  12.4

             Moyenne     Ecart-type     Coeff. de variation
Effet MASC = -5.9           5.7                 0.96
Effet JEUN = -5.0           7.5                 1.51
Effet MEDI = -1.5           4.1                 2.66
Effet BACS = 15.6           6.8                 0.44
Effet CSUP =  7.8           4.5                 0.58
```

Etre de sexe masculin, jeune ou d'âge médian fait diminuer le taux de fort lecteur, le fait d'être de classe supérieure l'augmente et encore plus d'avoir un niveau de diplôme égal ou supérieur au bac.

Pour les effets MASC, BACS et CSUP, la moyenne est plus grande en valeur absolue que l'écart-type, pour JEUN et surtout pour MEDI, elle lui est inférieure. Ces variations montrent que le modèle additif que nous supposons comprend des exceptions. Etudions l'une d'entre elle : par exemple pour l'effet moyen de sexe masculin on constate de grandes variations dans les sous-effets constitutifs et, si l'on prend les deux plus extrêmes, on a :

```
JEUN BACS CSUP -24.2  186
AGEE BACS CSUP   4.3   73
```

L'effet négatif signifie que dans la sous-population des jeunes diplômés de classe supérieure, le fait d'être de sexe masculin entraine une chute du taux de lecteurs, tandis que dans la sous-population des âgés également diplômés de classe supérieure, l'effet de sexe masculin est positif. En grossissant les traits, dans la classe supérieure diplômée, les jeunes garçons lisent moins que les jeunes filles et les vieux messieurs lisent plus que les vieilles dames. Pour retrouver ce résultat par un tri croisé ordinaire, il suffit d'isoler la sous-population des diplômés de classe supérieure jeunes ou âgés et de croiser l'âge et le sexe. On a l'analyse multivariée suivante :

```
Variable test AGE  modalité JEUN
Croisement question SEX  et question LEC
Le Khi-deux du tableau est de    14.0

  COL:  BCP  AUTR        BCP AUTR        BCP AUTR
```

```
MASC      13    79    92    14.1 85.9 100      -     +
FEMI      36    58    94    38.3 61.7 100      +     -
  TOT     49   137   186    26.3 73.7 100
```

On retrouve bien l'effet de sexe masculin avec une différence de 14,1 - 38,3 = -24,2

```
Variable test AGE  modalité AGEE
Croisement question SEX  et question LEC
Le Khi-deux du tableau est de     0.1

  COL:   BCP   AUTR         BCP AUTR         BCP AUTR
MASC      18    27    45    40.0 60.0 100      +     -
FEMI      10    18    28    35.7 64.3 100      -     +
  TOT     28    45    73    38.4 61.6 100
```

L'effet de sexe masculin se fait maintenant dans l'autre sens avec une différence de 40,0 - 35,7 = 4,3.[8].

Nous sommes en présence d'une interaction ayant entrainé une inversion de signes d'écart à l'indépendance lors d'une analyse multivariée. On s'aperçoit ainsi que ce sont les interactions qui perturbent le modèle additif : elles peuvent être de ce fait directement repérées dans le formalisme montré plus haut qui est simplement une analyse sur tri profond, que nous nommons *analyse tabulaire*.

Comparaison de modèles

L'exemple précédent semble cohérent avec un modèle additif dans la mesure où, lorsqu'il s'en écarte, il s'agit d'interactions qui peuvent être interprétées : il est donc raisonnable de voir ce que cette hypothèse donne en essayant une régression utilisant un modèle additif sur les données précédentes. A cette fin on met les données en codage disjonctif et, sous SAS, on construit le modèle de la régression logistique que l'on applique aussi à la régression linéaire :

```
data   ;
    INFILE 'c:\div\enq\prat\BMS22.cdg';
    input indiv$ 1-4 (
  MASC FEMI JEUN MEDI AGEE NDIP BACS CSUP CINF BCP AUTR) (4.);
    run;

PROC REG;
  model BCP =  MASC JEUN MEDI  BACS CSUP ;
run;

PROC LOGISTIC DESCENDING ;
  model BCP =  MASC JEUN MEDI  BACS CSUP ;
run;
```

La régression linéaire donne les résultats suivants :

```
Dependent Variable: BCP    Analysis of Variance
```

[8] Cependant le khi-deux nous indique que cette différence n'est pas significative alors que la précédente l'est tout à fait

```
                  Sum of          Mean
Source       DF        Squares            Square          F  Value
Prob>F
Model         5       42.74550          8.54910            65.437
0.0001
Error    4991    652.05978       0.13065
C Total 4996    694.80528
    Root MSE        0.36145     R-square       0.0615
    Dep Mean        0.16690     Adj R-sq       0.0606
    C.V.          216.56747
                   Parameter Estimates
                Parameter    Standard    T for H0:
Variable   DF    Estimate       Error     Parameter=0    Prob >
|T|
INTERCEP   1      0.144071    0.01028481    14.008       0.0001
MASC       1     -0.057449    0.01024252    -5.609       0.0001
JEUN       1     -0.064118    0.01507751    -4.253       0.0001
MEDI       1     -0.011629    0.0120127     -0.968       0.3331
BACS       1      0.151300    0.01230049    12.300       0.0001
CSUP       1      0.086526    0.01334305     6.485       0.0001
```

Comparons ces résultats et ceux de l'analyse tabulaire en traduisant les paramètres de la régression linéaire en pourcentage (au lieu des proportions)

```
Analyse tabulaire                Régression linéaire
Situation de
 référence = 12.4                       14.4
Effet MASC = -5.9                       -5.7
Effet JEUN = -5.0                       -6.4
Effet MEDI = -1.5                       -1.2
Effet BACS = 15.6                       15.1
Effet CSUP =  7.8                        8.7
```

Le moins que l'on puisse dire est que le choix du modèle linéaire donne un résultat dont la cohérence avec les données semble bonne. Seul l'effet MEDI est jugé non-significativement différent de zéro par le modèle linéaire : c'est celui qui dans l'analyse tabulaire avait le coefficient de variation le plus fort, c'est à dire celui où la dispersion était la plus forte et donc le moins cohérent avec l'unicité d'un effet.

Comparons maintenant avec les résultats de la régression logistique :

```
                    The LOGISTIC Procedure

Response Variable: BCP
Response Levels: 2
Number of Observations: 4997
Link Function: Logit
                          Response Profile
                          Ordered
                          Value      BCP      Count
                             1        1        834
                             2        0       4163
```

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	4508.662	4227.434	.
SC	4515.179	4266.533	.
-2 LOG L	4506.662	4215.434	291.229 with 5 DF (p=0.0001)
Score	.	.	307.423 with 5 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
ICPT	1	-1.8296	0.0806	515.4197	0.0001	.	.
MASC	1	-0.4496	0.0798	31.7321	0.0001	-0.123868	0.638
JEUN	1	-0.4866	0.1203	16.3683	0.0001	-0.107351	0.615
MEDI	1	-0.1184	0.0947	1.5631	0.2112	-0.032511	0.888
BACS	1	1.0379	0.0887	136.9652	0.0001	0.266767	2.823
CSUP	1	0.5458	0.0919	35.2456	0.0001	0.127524	1.726

Association of Predicted Probabilities and Observed Responses

Concordant = 63.7%	Somers' D = 0.341
Discordant = 29.6%	Gamma = 0.365
Tied = 6.7%	Tau-a = 0.095
(3471942 pairs)	c = 0.670

Comparons les trois résultats en transformant les paramètres de la régression logistique en effets marginaux en pourcentages[9]

Analyse tabulaire Situation de référence	Régression linéaire	Régression logistique
= 12.4	14.4	13.8
Effet MASC = -5.9	-5.7	-4.5
Effet JEUN = -5.0	-6.4	-4.8
Effet MEDI = -1.5	-1.2	-1.4
Effet BACS = 15.6	15.1	17.4
Effet CSUP = 7.8	8.7	7.9

Tous les résultats sont comparables et la régression logistique, comme la régression linéaire juge l'effet MEDI comme non différent de zéro. Avons-nous simplement l'embarras du choix et pouvons nous prendre le modèle linéaire simplement parce qu'il est plus proche de ce qu'on peut dire des données ? Une telle

[9] par la transformation 1 / 1 + exp -(alpha + Σbêta) qui donne la proportion de chaque situation dont on déduit les effets marginaux (qui sont mis en pourcentages)

éventualité soulève deux objections qu'il faut étudier : d'abord que le modèle linéaire ne peut pas s'appliquer à la situation, ensuite que le modèle logistique est meilleur.

Le modèle linéaire peut-il s'appliquer ?

La réponse classique est négative : pour Agresti (1990 : 84) "the linear probability model has a major structural defect. Probabilities must fall between 0 and 1, whereas linear functions take values over the entire real line" ; pour Gouriéroux (1989 : 9-11) "le modèle linéaire s'écrirait :

(1.1) $y_i = x_i b + u_i$ $i = 1,..., n.$

L'inadéquation d'une telle formulation peut facilement être mis en évidence par des arguments intuitifs et par des arguments mathématiques. Donnons-en quelques-uns : a) les deux membres de l'égalité (1.1) sont de nature différente : y_i est une variable qualitative et $x_i b + u_i$ est une variable quantitative, ce qui a évidemment peu de sens" (p.9). D'autres objections portent sur les hypothèse de normalité, les contraintes sur les $x_i b$ impossibles à respecter, les coefficients b impossibles à estimer (p.10-11).

La réponse classique a beau être négative, il faut bien se résoudre à la constatation que le modèle fonctionne toujours correctement, qu'il donne toujours des probabilités comprises entre 0 et 1, toujours proches et des analyses tabulaires et des résultats (mis en proportion) de la régression logistique, que la log-vraisemblance de la régression linéaire et de la régression logistique sont toujours extrêmement proches. Nous sommes donc devant un cas particulier dont il faut rendre compte.

On pourrait simplement dire *e pur si muove*, mais, sans chercher à résoudre analytiquement le problème, nous allons examiner le cas encore plus particulier d'un simple tableau croisé où il y a identité stricte entre les résultats de l'analyse tabulaire, de la régression linéaire et de la régression logistique. Tous les paramètres peuvent être calculés directement à partir des données de base.

En reprenant les données déjà étudiées, on croise le diplôme obtenu (nomenclature plus détaillée) avec le fait de beaucoup lire : on a le tableau croisé suivant :

Croisement question Diplôme et question Lecture
Le Khi-deux du tableau est de 285.9

COL:	BCP	AUTR	Tot	BCP	AUTR	Tot	
AUCN	201	1792	1993	10.1	89.9	100	Aucun diplôme ou CEP
BEPC	69	374	443	15.6	84.4	100	Bepc ou brevet
CAP	118	848	966	12.2	87.8	100	CAP et BEI
BAC	129	416	545	23.7	76.3	100	Bac
SUP	229	387	616	37.2	62.8	100	Etudes supérieures
ETUD	88	346	434	20.3	79.7	100	Etudes en cours
TOT	834	4163	4997	16.7	83.3	100	

On prend pour référence la catégorie intermédiaire du CAP qui nous donne le paramètre $alpha_{lin}$ de la régression linéaire en prenant simplement la proportion de fort lecteur c'est à dire 118 / 966 = 0,122153

Le paramètre $alpha_{log}$ de la régression logistique est le logit de cette proportion soit Ln(0,122153 / 1 - 0,122153) = - 1,972196

En régression linéaire, les paramètres de type bêta pour les cinq diplômes autres que la référence se calculent par simple différence entre la proportion de fort lecteur pour un diplôme et la proportion de référence. On a les résultats suivants :

	proportion de BCP	bêta	
AUCN	0.100853	−0.021300	
BEPC	0.155756	0.033603	
CAP	0.122153	0.000000	Référence
BAC	0.236697	0.114544	
SUP	0.371753	0.249600	
ETUD	0.202765	0.080612	

En régression logistique, les paramètres de type bêta pour les cinq diplômes autres que la référence se calculent par différence entre le logit d'un diplôme et celui de la référence. On a les résultats suivants :

	logit des prop.de BCP	bêta	
AUCN	−2.187783	−0.215587	
BEPC	−1.690149	0.282047	
CAP	−1.972196	0.000000	Référence
BAC	−1.170873	0.801323	
SUP	−0.524703	1.447493	
ETUD	−1.369102	0.603094	

En prenant ces paramètres, on calcule la probabilité p_i de chaque situation :
en régression linéaire par la formule : $p_i = alpha + bêta_k$
en régression logistique par la formule : $p_i = 1 / 1 + exp -(alpha + bêta_k)$

i indice les individus et k les diplômes : comme le tableau n'est qu'à deux dimensions, chaque individu n'est que dans une seule catégorie de diplôme et chaque individu n'est donc concerné que par un seul paramètre bêta.

Pour un individu donné, ces probabilités sont identiques par construction puisque les coefficients linéaires et logistiques se déduisent l'un de l'autre à partir des données de base. De ce fait la log-vraisemblance L_i est unique et est calculée pour chaque individu de la façon suivante (où y=1 si on est fort lecteur et 0 dans le cas contraire) :

$L_i = yi\ Ln(p_i) + (1 - Y_i)\ Ln\ (1 - p_i)$

Dans le tableau ci-dessous en codage disjonctif, à chaque ligne correspond une situation de données avec en tête l'effectif correspondant (on y retrouve tous les effectifs du tableau croisé). On donne la probabilité de chaque cas et sa log-vraisemblance (L_i individuelle multipliée par l'effectif correspondant). La sommation de toutes les log-vraisemblances individuelles correspond au même maximum de vraisemblance pour les deux types de régressions (puisque les probabilités sont identiques).

N	aucn	bepc	cap	bac	sup	etu	bcp	autr	P	L
201	1	0	0	0	0	0	1	0	0,1009	−461,11234
1792	1	0	0	0	0	0	0	1	0,1009	−190,50527
69	0	1	0	0	0	0	1	0	0,1558	−128,30306
374	0	1	0	0	0	0	0	1	0,1558	−63,32333

118	0	0	1	0	0	0	1	0	0,1222	−248,09275
848	0	0	1	0	0	0	0	1	0,1222	−110,47995
129	0	0	0	1	0	0	1	0	0,2367	−185,88570
416	0	0	0	1	0	0	0	1	0,2367	−112,36169
229	0	0	0	0	1	0	1	0	0,3718	−226,60137
387	0	0	0	0	1	0	0	1	0,3718	−179,88607
88	0	0	0	0	0	1	1	0	0,2028	−140,42227
346	0	0	0	0	0	1	0	1	0,2028	−78,40560
									Somme=	−2125,3794

La situation de cet exemple où il n'y a qu'un seul croisement manifeste bien les contraintes qui pèsent sur les paramètres et qui font que les probabilités dans le cas linéaire sont estimées à des valeurs comprises entre 0 et 1.

Dans le cas où la situation de référence correspond à plusieurs questions, la théorie reste à faire mais on devine que l'on s'écarte peu de la situation de l'exemple précédent.

Le modèle logistique est-il meilleur ? [10]

Apparu à l'époque contemporaine, l'usage du pourcentage semble dominer les utilisations chiffrées y compris dans la mesure des évolutions d'un phénomène[11]. Cependant se pose la question de savoir s'il faut noter les évolutions par une *différence* de pourcentage ou par un *rapport*. A partir d'un article de Combessie (1984), a eu lieu dans la *Revue française de sociologie* un débat qui s'est poursuivi sur plusieurs années et qui s'est semble-t-il terminé sur la victoire du modèle logistique avec les articles de Vallet (1988) et Euriat-Thélot (Euriat 1995). Dans ce dernier article, les auteurs notent dans une annexe (p.430) que le rapport logistique (appelé actuellement plus communément *odd ratio*) fait la synthèse entre l'étude de la différence et celle du rapport car il conduit aux mêmes conclusions que la *différence* quand les proportions sont extrêmes et aux mêmes conclusions que le r*apport* quand les proportions sont proches d'un demi.

Il va de soi qu'une différence de proportion n'a pas le même sens sur toute la gamme des proportions possibles. Par exemple, dans un phénomène de diffusion qui fonctionne par contagion, le début et la fin de croissance sont lents tandis que c'est dans la période intermédiaire que la croissance est la plus rapide : c'est la courbe en forme de sigma valable aussi bien pour l'acquisition d'un bien ménager que pour le taux des bacheliers et qui peut être modélisé par une fonction logistique (Cibois 1988).

Pour les paramètres d'une régression, le problème n'est pas celui d'une évolution mais celui de l'influence d'effets par rapport à une situation de référence qui fixe le domaine où vont se situer les effets. Un même effet aux environs de 10% n'a pas le même sens dans notre premier exemple où la situation de référence était de 62% que dans le second ou la situation de référence est de 12%. Cela n'a aucune importance puisque les effets ne sont comparés que pour une situation de référence donnée qui fixe en quelque sorte le spectre des variations. De plus, bien souvent, on

---

[10] Sur cette question, Louis-André Vallet a bien voulu réagir à une version précédente de cet article et je l'en remercie particulièrement.

[11] Par exemple sous l'Ancien régime, un impôt au taux de 5% est désigné par le chiffre du dénominateur et est appelé impôt du *vingtième*.

a vis-à-vis des effets une attitude assez distante quant à leur valeur précise. On s'intéresse plus au fait qu'ils soient positifs ou négatifs et à leur ordre de grandeur.

Il faut se souvenir que les discussions notées plus haut ont pour origine des débats sociaux sur l'évolution du système scolaire allant plus ou moins dans le sens de l'égalité, et que le débat devient pointilleux quand il est sous-tendu par des argumentations dans des domaines sensibles. En analyse des données, une régression est plus un outil de travail qui permet de synthétiser l'effet de variables, qu'un trébuchet qui permet de conclure à l'effet social d'une politique. On souhaite donc que l'instrument donne des résultats facilement interprétables et le modèle linéaire va dans ce sens.

Il faut noter enfin que ce souci de communication des résultats conduit souvent les utilisateurs de la régression logistique à traduire par une exponentiation leurs coefficients en pourcentages. L'exigence de lisibilité liée à un enjeu social fait que l'auteur conclut par un pourcentage afin d'être entendu.

En conclusion, que l'échelle logistique soit meilleure pour rendre compte d'évolutions ne fait pas que l'interprétation d'effets en pourcentages ne soit pas la solution la plus simple, cohérente avec l'emploi massif des pourcentages et des différences de pourcentages par les sociologues.

Conclusion

La procédure de recherche d'effets *toutes choses égales par ailleurs* appliquée à des tableaux croisés où l'on a pris une situation de référence, donne toujours des résultats proches tant de la régression linéaire que de la régression logistique : c'est ce fait massif qui emporte la conviction. Le choix du modèle linéaire s'impose du fait de sa simplicité et de sa cohérence avec les données. En ce qui concerne la ressemblance entre modèle linéaire et modèle logistique, des justifications théoriques sont encore à approfondir.

La permanence de la ressemblance de la régression linéaire et de la régression logistique (après exponentiation des paramètres) peut dès à présent être facilement testée par les chercheurs en remplaçant dans les logiciels usuels une méthode par une autre, sans rien changer des spécifications du modèle. Pour la comparaison avec l'analyse tabulaire on pourra soit demander le nouveau module ANATAB de Tri-deux fourni par l'auteur, soit, sous un logiciel standard précroiser toutes les variables explicatives et croiser la nouvelle variable avec la variable à expliquer. On obtient ainsi ce que nous avons appelé plus haut les *données de base* qui permettent de construire par soustraction les différents sous-effets dont la moyenne donnera chacun des effets.

Références

Agresti, Alan (1990). *Categorical Data Analysis*, New York, J.Wiley.

Berkson, J. (1944). "Application of the Logistic Function to Bio-Assay", *Journal of the American Statistical Association* 39 : 357-365.

Benzécri, Jean-Paul (1976). *L'analyse des données*, Paris, Dunod, 2 tomes, (1ère édition 1973).

Bouyer, J. (1991). "La régression logistique en épidémiologie", *Rev. Epidém. et Santé Publ.*, 39 : 79-87, 183-196.

Cibois, Philippe, Droesbeke, Jean-Jacques (1988). "La croissance du nombre des bacheliers est-elle modélisable et prévisible ? ", *Revue française de sociologie*, 29 (3), 425-445

Combessie, Jean-Claude (1984). "L'évolution comparée des inégalités : problèmes statistiques", *Revue française de sociologie*, 25 (2), 233-254

Cox, D.R. (1972). *Analyse des données binaires*, Paris, Dunod.

Edwards, A. W. F. (1992). *Likelihood*, Baltimore, The John Hopkins University Press, (1ère édition 1972).

Euriat Michel, Thélot Claude (1995). "Le recrutement social de l'élite scolaire en France", *Revue française de sociologie*, 36 (3), 403-438

Gouriéroux, Christian (1989). *Econométrie des variables qualitatives*, Paris, Economica, 2e édition.

Lollivier, S., Marpsat, M., Verger, D. (1996) "L'économétrie et l'étude des comportements", Paris, INSEE, série des documents de travail "Méthodologie statistique" n°9606, 78p.

Rouanet, Henry et Le Roux, Brigitte (1993). *Analyse des Données Multidimensionnelles*, Paris, Dunod.

Vallet, Louis-André (1988). "L'évolution de l'inégalité des chances devant l'enseignement", *Revue française de sociologie*, 29 (3), 395-423