Bulletin de Méthodologie Sociologique N°64 Juillet 2000 - p.54-64

Observation et modèle linéaire ou logistique : réponse à Aris et Hagenaars.

Philippe Cibois
Laboratoire Printemps,
Université de Versailles - St-Quentin
phcibois@wanadoo.fr

Résumé. La réponse apportée ici à Aris et Hagenaars (2000) insiste sur le fait que la mauvaise adéquation avec les données du modèle logistique ne porte que sur les effets marginaux en pourcentage. On plaide pour l'abandon de cet indicateur pour soit prendre les estimations des odds ratios en régression logistique, soit prendre les observations de l'analyse tabulaire (ou leur estimation par les effets marginaux en pourcentage de la régression linéaire). **Données qualitatives, Analyse tabulaire, Régression logistique, Régression linéaire.**

Les "remarques sur la comparaison entre les modèles linéaire et logit" d'Emmanuel Aris et Jacques Hagenaars (2000) veulent attirer l'attention sur le fait que le choix d'un modèle n'est pas dicté par l'adéquation aux données dans ce sens que le choix d'un modèle implique le choix d'un type de mesure de l'écart à l'indépendance qui permettra, le cas échéant, de prendre en compte les effets d'interaction.

Je vais essayer de montrer dans cette réponse que les réserves que l'on peut formuler sur l'utilisation du modèle logistique en régression sur variables nominales viennent des problèmes que posent l'interprétation des effets marginaux exprimés en pourcentage et que, pour ce paramètre, l'adéquation aux données est mauvaise.

Pour le montrer, je propose, en travaillant sur des données réelles, d'utiliser diverses méthodes et de comparer les interprétations qui peuvent en être proposées selon les choix qui peuvent être faits de la situation de référence.

Le choix de l'option latin en 4^{e}

En France actuellement, environ un quart de la classe d'âge choisit de faire du latin en option supplémentaire en 4^{e} : pour rendre compte de ce choix (présence de l'option latin LAT+ contre son absence LAT-), en utilisant les données du Panel 89 de l'Éducation nationale, on prend les indicateurs suivants :

1) l'origine sociale : supérieure (CSUP) si professions libérales, cadres supérieurs ou professions intermédiaires ; inférieure (CINF) si autre situation,

2) le gout pour la lecture de l'enfant : faible (LEC-) s'il n'a lu que 5 livres depuis le début de l'année ; fort (LEC+) s'il en a lu 6 et plus,

3) autre indicateur de lecture : l'élève déclare ou non si la lecture fait partie de ses loisirs favoris (FAV+ / FAV-),

4) niveau de mathématiques : bon niveau ou non (MAT+ / MAT-).

Voici les données de base où l'on croise toutes les situations des variables explicatives avec la variable à expliquer. Les données sont ordonnées par niveau décroissant du choix de l'option latin.

				LAT+	LAT-	Total	Prop.	
1	CSUP	LEC+	FAV+	MAT+	70	39	109	64.2	situation 1
2	CSUP	LEC+	FAV-	MAT+	24	17	41	58.5	
3	CSUP	LEC-	FAV+	MAT+	27	28	55	49.1	

4	CINF	LEC+	FAV+	MAT+	37	55	92	40.2	
5	CSUP	LEC-	FAV-	MAT+	34	54	88	38.6	
6	CSUP	LEC+	FAV+	MAT-	22	44	66	33.3	
7	CINF	LEC-	FAV+	MAT+	22	54	76	28.9	
8	CINF	LEC+	FAV-	MAT+	8	23	31	25.8	
9	CSUP	LEC-	FAV+	MAT-	11	39	50	22.0	
10	CINF	LEC-	FAV-	MAT+	30	120	150	20.0	situation 2
11	CSUP	LEC+	FAV-	MAT-	4	17	21	19.0	
12	CSUP	LEC-	FAV-	MAT-	14	92	106	13.2	
13	CINF	LEC+	FAV+	MAT-	12	174	186	6.5	
14	CINF	LEC-	FAV+	MAT-	12	203	215	5.6	
15	CINF	LEC-	FAV-	MAT-	11	485	496	2.2	situation 3
16	CINF	LEC+	FAV-	MAT-	1	82	83	1.2	
				Total	339	1526	1865	18.2	

Pour la suite, on va retenir trois situations de référence : celle qui correspond au plus haut niveau de l'option latin (Situation 1), une situation intermédiaire proche de la moyenne mais d'effectif suffisant (Situation 2), celle qui bien que correspondant à un très faible niveau d'option latin a cependant un effectif suffisant (Situation 3).

Nous allons faire, pour chacune de ces situations, trois analyses : analyse tabulaire (logiciel Trideux), régression linéaire (REG de SAS) et régression logistique (LOGISTIC de SAS). On donne pour l'analyse tabulaire la situation de référence observée et l'effet moyen ; pour les régressions, l'estimation de la situation de référence et des effets (en pourcentage dans les deux cas et en Odds-ratios pour la régression logistique).

Situation de référence n°1 : CSUP LEC+ FAV+ MAT+

	Analyse tabulaire	Reg.Lin	Reg.Log	Odds Ratio
Sit. Ref	64,2	56,8	68,3	
CINF	-18,2	-19,6	-31,6	0,269
LEC-	-5,3	-6,5	-11,0	0,623
FAV-	-6,7	-6,6	-14,0	0,552
MAT-	-24,7	-25,6	-42,7	0,160

Situation de référence n°2 : CINF LEC- FAV- MAT+

	Analyse tabulaire	Reg.Lin	Reg.Log	Odds Ratio
Sit. Ref	20,0	24,1	16,6	
CSUP	18,2	19,6	26,0	3,724
LEC+	5,3	6,5	7,6	1,606
FAV+	6,7	6,6	9,9	1,811
MAT-	-24,7	-25,6	-13,5	0,160

Situation de référence n°3 : CINF LEC- FAV- MAT-

	Analyse tabulaire	Reg.Lin	Reg.Log	Odds Ratio
Sit. Ref	2,2	-1,5	3,1	
CSUP	18,2	19,6	7,5	3,724
LEC+	5,3	6,5	1,8	1,606
FAV+	6,7	6,6	2,4	1,811
MAT+	24,7	25,6	13,5	6,267

Remarques préalables :

- *les paramètres négatifs en régression linéaire* : toutes les estimations des paramètres de régression sont significativement différentes de zéro sauf une, celle qui correspond à la situation de référence n°3, négative en régression linéaire. On voit sur cet exemple que la possibilité théorique d'un effet négatif en régression linéaire peut se rencontrer dans des situations réelles quand cette situation de référence est effectivement très proche de zéro. Dans ce cas la régression linéaire semble proposer une valeur qui peut être négative mais ici d'une manière non significative. Il est cohérent dans ce cas de considérer que la situation de référence est très proche de zéro.

- *l'indifférence de la situation de référence* : on sait que le choix de la situation de référence est indifférent dans ce type d'analyse. On voit en effet que pour chaque type d'analyse, il existe un invariant. Pour l'analyse tabulaire et la régression linéaire, c'est l'effet marginal qui s'inverse algébriquement quand on prend l'autre situation comme référence ; pour la régression logistique, c'est l'odds ratio (OR) dont l'inverse est pris quand on inverse la référence.

Interprétation :

Analyse tabulaire et régression linéaire : (ces deux méthodes, très similaires dans leurs résultats, reposant sur le même modèle additif, sont traitées de la même façon). L'interprétation prend acte de l'indifférence de la situation de référence qui peut être forte, moyenne ou quasiment nulle et tire la conclusion que ce qui explique d'abord le choix de l'option latin, c'est le fait que l'élève soit un bon élève, ce qu'indique son niveau en mathématique. L'effet des mathématiques est de l'ordre de 25%, quelque soit la méthode, en plus ou en moins selon la situation de référence choisie.

L'effet d'origine sociale n'arrive qu'en second avec une valeur proche de 20%. Si le latin est recherché pour des raisons sociales, ces raisons n'arrivent qu'en second lieu. Enfin l'aspect littéraire du choix du latin n'arrive qu'en troisième position : les deux indicateurs d'un choix en faveur de la lecture sont équivalents et de l'ordre de 6%.

On pourrait critiquer la première interprétation relative à l'indicateur "mathématiques" et mettre en valeur le fait que le latin, de par sa structure logique, de par la rigueur nécessaire pour sa pratique, est choisi par ceux dont le niveau mathématique manifeste leur gout pour la rigueur intellectuelle. Une analyse

complémentaire (Cibois 1996) a montré que dans la sous-population des bons élèves, le choix du latin était fait soit pour des raisons sociales, soit du fait d'un choix littéraire. Sur les origines de l'argumentaire en faveur du latin cf Cibois (2000).

Régression logistique : on doit distinguer l'interprétation rigoureuse, qui n'utilise que les OR qui, à une inversion près sont invariants, d'une interprétation qui tenterait d'utiliser les effets marginaux en pourcentage.

1) si l'on utilise que les OR, en situation 3 on doit dire que les chances de faire l'option latin sont multipliées par 6,3 par rapport aux chances associées à la situation de référence, alors qu'elles sont multipliées par 3,7 pour l'origine sociale et seulement par 1,6 à 1,8 pour les deux indicateurs de choix littéraire. On retrouve ici facilement les résultats précédents.

Nous entendons "chances" comme traduction de "odds" (p / 1-p) et "rapport des chances" comme traduction possible de "odds-ratio" ; "chances" sera remplacé par "risques" en fonction du contexte.

En situation 1, on arrive à la même interprétation mais avec un langage plus complexe. En effet il faut comparer des multiplicateurs inférieurs à un : on peut soit prendre le même langage que précédemment en employant les inverses et en parlant de division plutôt que de multiplication. On dira que les chances sont divisées par 3,7 ou 6,3 si l'on est de classe inférieure ou si on n'a pas un bon niveau de math. On peut aussi suivre de plus près le nombre affiché et dire que les chances quand on est de classe inférieure sont, par rapport à la situation de référence dans le rapport de 27 à 100 tandis qu'elles sont dans le rapport de 16 à 100 pour le faible niveau de math. La baisse des chances est beaucoup plus forte dans ce dernier cas.

Enfin, en situation 2, comme il y a des OR plus grands que l'unité et une autre inférieur, pour rendre l'intensité des effets comparables il faudra dire que dans un cas on divise ses chances par 6,3 (si l'on n'a pas un bon niveau en math) ou qu'on les multiplie par 3,7 (si l'on est d'origine sociale supérieure) toujours par rapport aux chances de la situation de référence.

Les OR (ou rapports des chances) ont un sens rigoureux mais il suffit de lire les publications qui utilisent la régression logistique dans le domaine des sciences sociales pour se rendre compte qu'ils sont rarement utilisés car ils supposent de l'utilisateur une compréhension de l'outil rarement effective. De ce fait, ils sont souvent traduits en termes d'effets marginaux en pourcentage.

2) examinons donc l'interprétation qui peut être faite des effets marginaux en pourcentage. En comparant les situations 1 et 3 on voit que, *à des niveaux d'effets différents*, on peut faire les mêmes interprétations que précédemment : l'effet du niveau mathématique est plus fort que celui de l'origine sociale, lui-même plus fort que les deux effets "littéraires" eux-mêmes équivalents. On tiendra qu'il est normal que les effets, identiques en OR, soient différents en effets marginaux en pourcentages puisqu'ils ne se situent pas dans les mêmes niveaux de l'échelle des proportions. En situation 1 on est proche de 68% tandis qu'on est assez proche de zéro en situation 3.

Plusieurs remarques s'imposent cependant :

- en situation 1, on en arrive au cas absurde où la somme des effets fait baisser la proportion à -31% : on sait que l'on n'a pas le droit de faire cette somme et qu'il faut cumuler les produits d'OR ce qui permet de retrouver alors la situation de

référence n°3 égale à 3%. On se trouve dans la situation paradoxale, mal vécue en général par les commentateurs, où l'on donne des effets en pourcentage en plus ou en moins, avec interdiction d'en faire la somme. On est pris dans la contradiction entre donner un indicateur correct non compris et un indicateur trompeur mais compréhensible ;

- enfin, en situation 2, l'usage de l'indicateur "effet marginal en pourcentage" devient impossible : il faudrait faire comprendre à l'utilisateur que l'effet du faible niveau de mathématiques, comme il est négatif et fait baisser la situation de référence et donc se situe donc dans la sphère basse des proportions, est en fait *plus important* (bien que plus faible en valeur absolue) que l'effet de la classe supérieure qui n'est apparemment plus important que parce qu'il se dirige vers les niveaux élevés des proportions. Ce serait tendre un piège à l'utilisateur.

Quelle méthode prendre ?

La conclusion des remarques précédentes est que si l'on utilise la régression logistique, il est dangereux de présenter les effets marginaux en pourcentage et que seuls les rapports de chances permettent une interprétation rigoureuse.

Cependant, les effets marginaux en pourcentage *peuvent être observés* et l'on va voir qu'ils ne ressemblent pas aux effets issus de la régression logistique mais à ceux issus de la régression linéaire. Pour le montrer, essayons d'*observer* et non d'*estimer* l'effet du faible niveau en mathématiques pour la situation de référence n°1.

Nous partons du principe que puisque nous réfléchissons "toutes choses égales par ailleurs", on peut observer la valeur de l'effet pour chacune des situations où cet effet s'exerce. Il n'y a que huit cas de figures possibles, ce qui permet l'examen de chacun. Pour trouver ces différentes situations, il faut, puisque nous croisons le niveau de mathématique et le choix de l'option latin, faire varier les trois autres situations : origine sociale, niveau de lecture, lecture favorite ou non. Comme ces situations sont dichotomiques, leur composition engendre donc huit cas de figures :

Commençons par les situations extrêmes : soit le croisement des math et du latin dans la sous-population des individus de classe supérieure, lisant beaucoup et pour lesquels la lecture est une distraction favorite (CSUP, LEC+, FAV+) : ils sont 175. Le croisement est le suivant (on remarque que les deux lignes du tableau correspondent aux lignes 1 et 6 des données d'origine) :

	LAT+	LAT-	Tot	LAT+	LAT-	
MAT+	70	39	109	64.2	35.8	100
MAT-	22	44	66	33.3	66.7	100
Tot	92	83	175	52.6	47.4	100

On voit que le fait de passer du niveau fort au niveau faible des mathématiques fait passer le pourcentage de l'option latin de 64,2 à 33,3 soit un effet marginal de -30,9

Prenons maintenant l'autre situation extrême (CINF, LEC-, FAV-) correspondant aux lignes 10 et 15 des données d'origine.

	LAT+	LAT-	Tot	LAT+	LAT-	
MAT+	30	120	150	20.0	80.0	100
MAT-	11	485	496	2.2	97.8	100
Tot	41	605	646	6.3	93.7	100

L'effet marginal fait passer le pourcentage de 20,0 à 2,2 soit un effet de -17,8.

D'une manière analogue on calcule les 6 autres effets : on a, par ordre décroissant d'effet, la récapitulation suivante (on donne pour chaque situation : l'effet, la population associée et les deux lignes du tableau de données dont la soustraction engendre l'effet) :

```
CSUP LEC+ FAV- -39.5   62 lignes 2 et 11
CINF LEC+ FAV+ -33.8  278 lignes 4 et 13
CSUP LEC+ FAV+ -30.9  175 lignes 1 et 6
CSUP LEC- FAV+ -27.1  105 lignes 3 et 9
CSUP LEC- FAV- -25.4  194 lignes 5 et 12
CINF LEC+ FAV- -24.6  114 lignes 8 et 16
CINF LEC- FAV+ -23.4  291 lignes 7 et 14
CINF LEC- FAV- -17.8  646 lignes 10 et 15
```

Ces huit effets sont de même ordre : quelque soit la situation, le fait d'avoir un faible niveau de mathématiques fait baisser le choix de l'option latin d'environ 20 à 30% ce qui fait qu'on peut résumer cette observation en prenant la valeur moyenne pondérée de -24,7%.

Comme cette valeur, qui est donc celle de l'analyse tabulaire et qui est proche de la valeur estimée par la régression linéaire, est le fruit d'une observation qui utilise un outil utilisé par tous les sociologues, les différences de pourcentages dans un tableau croisé, j'arrive à la conclusion qu'il faut préférer cette observation (ou les estimations de la régression linéaire qui en sont proches) aux estimations des effets en pourcentages de la régression logistique. On a là un outil familier qui nous permet de voir que, quelque soit la situation contrôlée, c'est à dire toutes choses égales par ailleurs, l'effet des mathématiques se situe toujours aux alentours de 25% en valeur absolue.

Ce qui est important de voir, c'est que l'idéal scientifique exprimé par la revendication "toutes choses égales par ailleurs" est ici facilement opérationalisable par de simples tris croisés sur des sous-populations.

Valeur du modèle logistique

Cependant, il ne faut pas nier la qualité du modèle logistique : le débat passé de la *Revue française de sociologie*, de Combessie (1984) à Vallet (1988) a permis de comprendre que pour suivre l'évolution d'un phénomène de diffusion dans le temps, la loi logistique permettait de rendre compte des débuts et des fins à faible évolution et du milieu en croissance plus forte. De même, ce modèle nous aide à bien prendre conscience que les variations de proportions n'ont pas la même intensité tout au long de l'échelle. Ceci est un acquis important dont on constate d'ailleurs qu'il se vérifie dans le dernier tableau donné : si l'on compare la valeur de l'effet avec la valeur correspondant au pourcentage d'option latin dans le cas du haut niveau de mathématiques on a une décroissance des effets en fonction du point de départ qui va bien dans le sens de la variation de l'échelle des proportions :

```
               Effet %LAT+ si MAT+
CSUP LEC+ FAV- -39.5  58.5
CINF LEC+ FAV+ -33.8  40.2
CSUP LEC+ FAV+ -30.9  64.2
CSUP LEC- FAV+ -27.1  49.1
CSUP LEC- FAV- -25.4  38.6
CINF LEC+ FAV- -24.6  25.8
```

```
CINF LEC- FAV+ -23.4  28.9
CINF LEC- FAV- -17.8  20.0
```

Conclusion

On voit sur cet exemple que les inconvénients du modèle logistique en régression viennent du fait que les effets marginaux en pourcentage 1) posent des problèmes d'interprétation nombreux 2) sont trop différents des observations qu'ils estiment.

Ce n'est pas parce que le modèle logistique est complexe qu'il doit être éliminé mais inversement ce n'est pas parce qu'il est complexe qu'il doit être préféré. S'il n'y avait qu'un problème de pédagogie, il serait vite résolu : mon expérience d'enseignant me montre que des étudiants de maitrise jonglent rapidement avec les OR (pas tous cependant) et comprennent la méthode du maximum de vraisemblance si on ne se contente pas de la réduire à une formule mais qu'on en revient aux intuitions de Fisher (en étudiant le rapport de vraisemblance). Inversement, l'analyse tabulaire qui permet de comprendre en profondeur la démarche "toutes choses égales par ailleurs" est plus simple à assimiler que la méthode du maximum de vraisemblance et les OR mais ce n'est pas parce qu'elle est plus simple (trop simple ?) qu'elle ne doit pas être prise en compte.

Enfin, il faut souligner que nous sommes ici dans une revue de méthodologie sociologique : c'est en fonction d'un public de sociologues que l'usage de l'analyse multivariée, des tableaux croisés, des différences de pourcentage, d'un modèle simple additif, est cohérent. Que l'usage des OR du modèle de la régression logistique soit cohérent avec la pratique des chercheurs en épidémiologie est un problème qui ne nous concerne pas.

Je voudrais terminer en rappelant le principe initial de Tukey dans la préface d'*Exploratory Data Analysis* (1977) : "It is important to understand what you **can do** before you learn to measure how **well** you seem to have **done** it". Avant de chercher à qualifier la valeur d'un modèle, il faut regarder ce qu'on peut faire avec les données, il faut se laisser guider par l'observation : c'est ce que j'essaye de faire avec l'analyse tabulaire.

Références

Aris, Emmanuel et Hagenaars, Jacques (2000). Remarques sur la comparaison entre modèles linéaire et logit", *Bulletin de méthodologie sociologique*, avril n°66 : 5-12.

Cibois, Philippe (1996). "Le choix de l'option latin au Collège" , *Education et Formations*, n°48, décembre : 39-51.

Cibois, Philippe (1999). "Modèle linéaire contre modèle logistique en régression sur données quantitatives", *Bulletin de méthodologie sociologique*, octobre n°64 : 5-24.

Cibois, Philippe (2000). "La question du latin : des critiques du XVIII^{e} siècle au *revival* du XIX^{e}", *L'Information littéraire*, 52 (1), janvier-mars : 7-28.

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

Tukey, John W. (1977). *Exploratory Data Analysis*, Reading Mass. : Addison-Wesley.

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