Utiliser le modèle log-linéaire pour mettre au jour la structure du lien entre les deux variables d'un tableau de contingence : un exemple d'application à la mobilité sociale

Louis-André Vallet (CNRS)

Laboratoire de Sociologie Quantitative,
CREST, UMR 2773 CNRS & INSEE

- Qu'est-ce que le modèle log-linéaire d'un tableau de contingence (à 2, 3, 4 ou k dimensions) ?

  C'est en fait un modèle multiplicatif pour les effectifs qui figurent dans les différentes cellules de ce tableau : chaque effectif peut être conçu comme un produit de paramètres.

- Un tel modèle est utile pour décrire la force et la structure des associations statistiques entre les variables croisées.

  En particulier, son intérêt est que les associations statistiques y sont décrites au moyen de la statistique du *odds ratio*, c'est-à-dire indépendamment des distributions marginales.

- La régression logistique d'une variable dépendante qualitative sur un ensemble de variables qualitatives considérées comme explicatives constitue un cas particulier du modèle log-linéaire du tableau de contingence croisant toutes ces variables.

  Par exemple, la régression logistique de E sur A, B, C et D (sans interaction entre ces variables dans leur effet sur la variable dépendante) correspond au modèle log-linéaire non saturé

  {ABCD, AE, BE, CE, DE}

  En d'autres termes, on sature la relation entre toutes les variables explicatives et on permet une association partielle entre chacune d'elles et la variable dépendante.

- Pour un document de cours général, voir :

  http://www.crest.fr/pageperso/lsq/vallet/vallet.htm

- Dans cette communication, on va utiliser le modèle log-linéaire pour résoudre un problème qui pourrait sembler apparemment très simple.

  Il s'agit de mettre au jour la structure de l'association statistique entre les deux variables d'un tableau de contingence à deux dimensions.

- L'exemple retenu est celui d'une table de mobilité sociale (7 lignes, 7 colonnes) issue de l'enquête Formation-Qualification Professionnelle de 1985.

  Pour les femmes françaises, actives occupées, âgées de 35 à 59 ans en 1985 (N=5178), on croise la catégorie sociale de leur père avec leur catégorie sociale personnelle, dans une nomenclature en sept postes.

- *Le problème à résoudre est le suivant*. Dans un tableau de contingence à deux dimensions, les effectifs des différentes cellules *confondent*, c'est-à-dire reflètent à la fois, le lien ou association statistique entre les deux variables croisées et l'importance relative de leurs différentes modalités.

  Dans notre exemple, 20,6% des femmes sont employées filles d'un contremaître ou d'un ouvrier, alors que 0,2% sont chefs d'entreprise ou professions libérales filles d'un homme de cette catégorie.

  La différence extrême entre ces deux proportions reflète bien sûr la forte importance relative des ouvriers dans la structure sociale des pères comme la forte importance relative des employées dans la structure sociale des filles.

  Il est ici impossible de savoir si la propension à l'hérédité sociale dans les classes supérieures est plus, aussi ou moins élevée que la tendance à la mobilité des filles d'ouvrier vers la catégorie des employées.

Le problème est le même pour les pourcentages en ligne (point de vue de la destinée) et les pourcentages en colonne (point de vue du recrutement).

- Bref, les effectifs ou proportions de la table de mobilité ont un intérêt évident pour *décrire la mobilité observée*.

- Mais, parce qu'ils dépendent à la fois de l'effet d'association entre les variables *et* des effets de marges, *ils ne peuvent nous procurer une vision directe de cette association, c'est-à-dire des proximités ou distances entre catégories sociales d'une génération à l'autre*.

- Or, en utilisant la gamme des modèles multiplicatifs (ou log-linéaires) situés entre le modèle d'indépendance statistique et le modèle saturé, on va pouvoir séparer les tendances intrinsèques à l'immobilité et à la mobilité des effets de marges.

- En d'autres termes, la modélisation des paramètres d'association va nous permettre de *décrire la fluidité sociale*.

Mobilité « parfaite » ou indépendance statistique

$$m_{ij} = \alpha\beta_i\gamma_j\delta_{ij} \qquad \delta_{ij} = 1 \quad \forall i, \forall j$$

$$Log\ m_{ij} = a + b_i + c_j + d_{ij} \qquad d_{ij} = 0 \quad \forall i, \forall j$$

Il est naturel de débuter l'analyse par ce modèle très simple qui exprime l'hypothèse d'absence d'association entre les variables.

Chaque effectif attendu est le produit de trois paramètres :

- une constante dont la présence va garantir l'ajustement de l'effectif total ;
- un paramètre ligne dont la présence va garantir l'ajustement de la structure sociale des pères ;
- un paramètre colonne dont la présence va garantir l'ajustement de la structure sociale des filles.

Il y a aussi la constante 1 qui reflète une « densité » *uniforme* sur toute la table de mobilité. En d'autres termes, sous ce modèle, *toutes les catégories sociales sont également proches ou distantes les unes des autres*.

En effet, sous un tel modèle, la concurrence entre les femmes de deux origines distinctes (i et i') pour atteindre (ou éviter) l'une plutôt que l'autre de deux positions sociales (j et j') est parfaite car :

$$\frac{m_{ij} \Big/ m_{ij'}}{m_{i'j} \Big/ m_{i'j'}} = 1$$

On ne s'étonnera donc pas qu'un tel modèle doive être rejeté.

En effet, le khi-deux du rapport de vraisemblance vaut presque 1550 pour 36 degrés de liberté !

On ne peut donc considérer que toutes les catégories sociales sont également proches ou distantes les unes des autres.

Ou encore, on ne peut considérer qu'indépendamment des effets de structure, tous les trajets de mobilité et d'immobilité sont également vraisemblables.

Les chausse-trappes du rapport de mobilité

$$R_{ij} = \frac{n_{ij}}{m_{ij}} = \frac{n_{ij} n_{..}}{n_{i.} n_{.j}}$$

Formée comme le rapport d'un effectif observé à l'effectif attendu sous la mobilité parfaite, cette statistique a été utilisée à partir des années cinquante pour résoudre notre problème et décrire les tendances intrinsèques à l'immobilité et à la mobilité dans chaque catégorie sociale.

Or, *cet usage pose problème* car, pour plusieurs raisons, le rapport de mobilité ne peut fournir une image exacte de l'association statistique, nette des effets de structure.

En particulier, s'il vaut 0 au minimum et 1 dans le cas de l'indépendance statistique, sa valeur maximale dans une cellule donnée dépend des marges correspondantes du tableau. En comparant sa valeur dans deux cellules diagonales, on ne peut donc savoir dans quelle catégorie sociale la tendance intrinsèque à l'immobilité est la plus forte.

Plus généralement, parce qu'il est fondé sur un modèle qui ne s'ajuste pas correctement, le rapport de mobilité ne peut *à la fois* mesurer l'écart des données à l'hypothèse de mobilité parfaite et décrire l'association statistique net des effets de structure.

Effectuant la première tâche, il ne peut accomplir correctement la seconde.

Il nous faut donc rechercher un modèle plus adéquat.

C'est en premier lieu parce qu'elle sous-estime l'ampleur de l'immobilité dans les catégories sociales que l'hypothèse de mobilité parfaite est inacceptable.

Cela suggère donc le modèle suivant qui, en un sens, est « le plus proche possible » de la mobilité parfaite…

Hérédité sociale uniforme

$$m_{ij} = \alpha\beta_i\gamma_j\delta_{ij} \qquad \delta_{ij} = \delta \; si \; i = j, \delta_{ij} = 1 \text{ sinon}$$

$$Log\; m_{ij} = a + b_i + c_j + d_{ij} \qquad d_{ij} = d \; si \; i = j, d_{ij} = 0 \text{ sinon}$$

Un tel modèle suppose que :

- pour les femmes qui quittent leur milieu d'origine, le mouvement vers telle ou telle position est toujours gouverné par l'hypothèse de mobilité parfaite ;
- il existe par ailleurs une tendance à hériter de la position paternelle (si *d* est strictement positif), tendance dont la force est *identique* dans chaque catégorie sociale.

Ainsi, le modèle postule que le fait d'être née dans une classe donnée procure à une femme un « avantage » pour demeurer dans la même position et, plus précisément, *le même* « avantage » dans chacune des classes.

Sur les données analysées, l'estimateur du paramètre *d* vaut 1,005 (0,040). Parmi les femmes qui exercent un emploi en 1985, il existe donc bien une tendance à conserver la position sociale paternelle.

Cependant, même s'il élimine près de 40% de la valeur du khi-deux obtenue sous l'hypothèse de mobilité parfaite, le modèle d'hérédité sociale uniforme n'est pas admissible : la statistique de test vaut plus de 930 pour 35 degrés de liberté !

Il faut donc conclure que :

- la tendance à conserver la position paternelle ne revêt pas la même force dans les différentes catégories sociales ;
- et/ou l'existence de distances entre les catégories contrarie la mobilité (au sens strict) des pères aux filles.

Relaxer la contrainte d'uniformité de la tendance à l'hérédité sociale forme donc la prochaine étape…

Mobilité « quasi parfaite » ou quasi-indépendance statistique

$$m_{ij} = \alpha\beta_i\gamma_j\delta_{ij} \qquad \delta_{ij} = \delta_i \; si \; i = j, \delta_{ij} = 1 \text{sinon}$$

$$Log\; m_{ij} = a + b_i + c_j + d_{ij} \qquad d_{ij} = d_i \; si \; i = j, d_{ij} = 0 \text{sinon}$$

On admet ici que *la tendance à l'hérédité sociale diffère d'une catégorie à une autre*. L'interprétation sociologique est donc la suivante : être née dans une classe sociale donnée procure un « avantage » pour y demeurer, mais l'importance de celui-ci varie selon la classe.

De nouveau, un tel modèle n'est pas acceptable car la statistique de test vaut 451 pour 29 degrés de liberté. Cependant, par rapport au modèle précédent, l'amélioration est substantielle.

C'est donc que les différences entre catégories sociales dans la propension à l'immobilité sont hautement significatives. Il convient alors d'examiner les estimations des paramètres correspondants…

$d_1 = 1{,}813\ (0{,}146)$ $d_2 = 1{,}753\ (0{,}329)$ $d_3 = 0{,}611\ (0{,}115)$ $d_4 = 0{,}682\ (0{,}115)$

$d_5 = -0{,}027\ (0{,}099)$ $d_6 = 0{,}657\ (0{,}090)$ $d_7 = 3{,}060\ (0{,}131)$

Ainsi, des pères aux filles, la tendance intrinsèque à l'hérédité sociale est :

- la plus forte chez les agriculteurs exploitants (7) ;
- moins élevée et du même ordre dans les fractions salariée et indépendante des catégories supérieures (cadres et professions intellectuelles supérieures (1), chefs d'entreprise et professions libérales (2)) ;
- plus faible, mais encore très significative parmi les professions intermédiaires (3), artisans et commerçants (4), contremaîtres et ouvriers (6) ;
- inexistante dans la catégorie des employés (5).

On vérifie en outre qu'un modèle de mobilité quasi parfaite muni des contraintes suivantes ne détériore pas significativement l'ajustement :

$$d_1 = d_2 \qquad d_3 = d_4 = d_6 \qquad d_5 = 0$$

Pour synthétiser, aucun des modèles envisagés jusqu'à présent n'a permis une analyse satisfaisante du lien entre milieu d'origine et position occupée, net des effets de structure !

Mais les hypothèses émises sur cette association ont été très simplificatrices : on a toujours supposé que la mobilité (au sens strict) était sous-tendue par une fluidité parfaite ! *Il reste donc à mettre au jour les « distances sociales » qui préforment la mobilité d'une catégorie d'origine à une catégorie de destination.*

Par ailleurs, *tous les modèles précédents affectent des niveaux de densité aux cellules de la table de mobilité* :

- un seul niveau pour le modèle de mobilité parfaite ;
- deux niveaux pour celui d'hérédité sociale uniforme ;
- huit niveaux pour le modèle de mobilité quasi parfaite ;
- quatre seulement si l'on incorpore nos contraintes finales.

Il reste alors à généraliser cette approche…

Estimer les rapports de densité dans la table de mobilité :
le modèle topologique ou structural de Hauser (1978, 1980)

Avec une partition des cellules *(i, j)* de la table en *K* sous-ensembles :

$$m_{ij} = \alpha\beta_i\gamma_j\delta_{ij} \qquad Log\ m_{ij} = a + b_i + c_j + d_{ij}$$

$$\delta_{ij} = \delta_k \qquad d_{ij} = d_k$$

si la cellule *(i, j)* appartient au sous-ensemble *k*

« Les effectifs attendus sont le produit d'un effet d'ensemble (α), d'un effet ligne (β_i), d'un effet colonne (γ_j) et d'un effet d'interaction (δ_{ij}). Les paramètres de ligne et de colonne correspondent au concept de prévalence. Ils reflètent l'offre et la demande professionnelles, les processus de remplacement démographique et les conditions économiques et technologiques passées et présentes. Les cellules (i, j) sont assignées à K sous-ensembles mutuellement exclusifs et exhaustifs, et celles de chaque sous-ensemble partagent un même paramètre d'interaction (δ_k). Ainsi, en dehors des effets d'ensemble, de ligne et de colonne, chaque effectif attendu n'est déterminé que par un paramètre d'interaction qui reflète la densité de mobilité ou d'immobilité dans cette cellule relativement à celle des autres cellules de la table. Les paramètres d'interaction du modèle correspondent directement au concept de densité commune des observations (White, 1963, p. 26), et ils peuvent être interprétés comme des indices de la distance sociale entre les catégories de ligne et de colonne du tableau croisé (comparer à Rogoff, 1953, pp. 31-32). »

Construire le modèle

De nombreuses décisions doivent cependant être prises avant de procéder à l'estimation statistique.

La première concerne le nombre de niveaux de densité. Hauser a souligné que l'interprétation du modèle devient malaisée lorsque ce nombre est grand. On retiendra donc *7 niveaux de densité, numérotés de* I *(supposé être le plus fort) à* VII *(le plus faible)*.

La seconde concerne l'affectation d'un niveau de densité à chacune des 49 cellules de la table de mobilité. Plutôt que de procéder de façon exploratoire (Hauser, 1980), on choisit de se fonder sur un ensemble de propositions théoriques pour être en mesure de tester leur pertinence a posteriori.

Une première hypothèse est que des forces puissantes concourent à l'immobilité sociale des femmes, d'où des densités élevées dans les cellules diagonales et en suivant notre dernier modèle : I (agriculteurs exploitants), II (cadres, chefs d'entreprise et professions libérales), IV (professions intermédiaires, artisans et commerçants, contremaîtres et ouvriers).

En ce qui concerne les trajectoires de mobilité, nous décidons de contraindre les niveaux de densité à respecter la symétrie par rapport à la diagonale principale. En d'autres termes, *nous postulons que les propensions à la mobilité de i vers j et de j vers i sont égales !*

En effet, l'existence de telles asymétries dans la structure sociale devrait être considérée comme un résultat sociologique important. Il convient donc de laisser à l'estimation statistique le soin d'établir si tel est le cas !

Une fois cette décision prise, il reste à allouer des niveaux de densité aux cellules du triangle supérieur de la table de mobilité féminine.

Cela est possible à partir d'un raisonnement sociologique qui peut être illustré à propos des filles de cadre et profession intellectuelle supérieure (cf. texte pp. 9-10).

On parvient finalement au modèle postulé suivant…

Matrice symétrique des niveaux de densité caractéristique de notre modèle de Hauser (Modèle de base)

	1	2	3	4	5	6	7
1 – Cadres, prof. int. Supérieures	II	III	IV	V	VI	VII	VII
2 – Chefs d'entr., Prof. libérales	III	II	IV	IV	VI	VII	VII
3 – Professions intermédiaires	IV	IV	IV	V	V	VI	VII
4 – Artisans, Commerçants	V	IV	V	IV	V	VI	VI
5 – Employés	VI	VI	V	V	V	V	VI
6 – Contremaîtres, Ouvriers	VII	VII	VI	VI	V	IV	V
7 – Agriculteurs exploitants	VII	VII	VII	VI	VI	V	I

Les lignes et les colonnes de la matrice correspondent respectivement à l'origine sociale (PCS du père) et à la position sociale (PCS de la fille). Parmi les niveaux de densité, I est supposé être le plus fort, VII le plus faible.

Pour ce modèle, la statistique de test vaut 81,70 pour 30 degrés de liberté. Il est donc, au sens strict, rejeté. Cependant, il s'agit du premier modèle qui doit être préféré au modèle saturé et il ne classe de manière erronée que 3,5% des femmes.

Surtout, les estimations des niveaux de densité s'ordonnent bien comme on le supposait au départ : de I, le plus fort, à VII, le plus faible !

Enfin, le même modèle estimé sur la table analogue de 1977 fournit une qualité d'ajustement et des estimations très proches.

Améliorer le modèle

À partir d'un examen des résidus, sept modifications qui, au total, mettent en jeu neuf cellules suffisent à obtenir une description satisfaisante de la fluidité sociale. Par exemple :

(Modification 1) Notre modèle de base sous-estime la mobilité des filles de gros indépendant dans le monde agricole. Nous affectons donc la densité V à la cellule (2, 7), mais n'accordons pas à cette modification de deux niveaux une importance majeure : l'effectif observé est faible, donc assez incertain, et une transformation analogue ne s'impose pas en 1977.

(Modification 2) Pour les filles d'artisan et commerçant, le mouvement vers la catégorie des employées est plus faible que nous ne le supposions. Nous décidons d'appliquer la même densité que pour la destinée ouvrière ou agricole (VI), transformation instructive car elle suggère l'existence d'une « rigidité à la baisse » dans la petite bourgeoisie indépendante.

(Modification 3) Elle est particulièrement importante : les modèles précédents surestiment les échanges, dans les deux directions, entre la catégorie des cadres et celle des agriculteurs exploitants alors qu'on supposait déjà ceux-ci très faibles (densité VII). *Créer pour les cellules (1, 7) et (7, 1) un niveau minimal de densité (VIII)* améliore alors sensiblement l'ajustement.

Le modèle final auquel on parvient ainsi fournit une statistique de test égale à 36,38 pour 30 degrés de liberté. *Le tableau suivant fournit les estimateurs des paramètres log-linéaires d_k et les rapports entre densités qui s'en déduisent.* À partir de là, il est possible de calculer la valeur des *odds ratios* impliqués.

Densité	Valeur du paramètre	
I	4,171	(0,263)
II	3,153	(0,250)
III	2,297	(0,243)
IV	1,714	(0,241)
V	1,279	(0,239)
VI	0,761	(0,238)
VII	0,000	

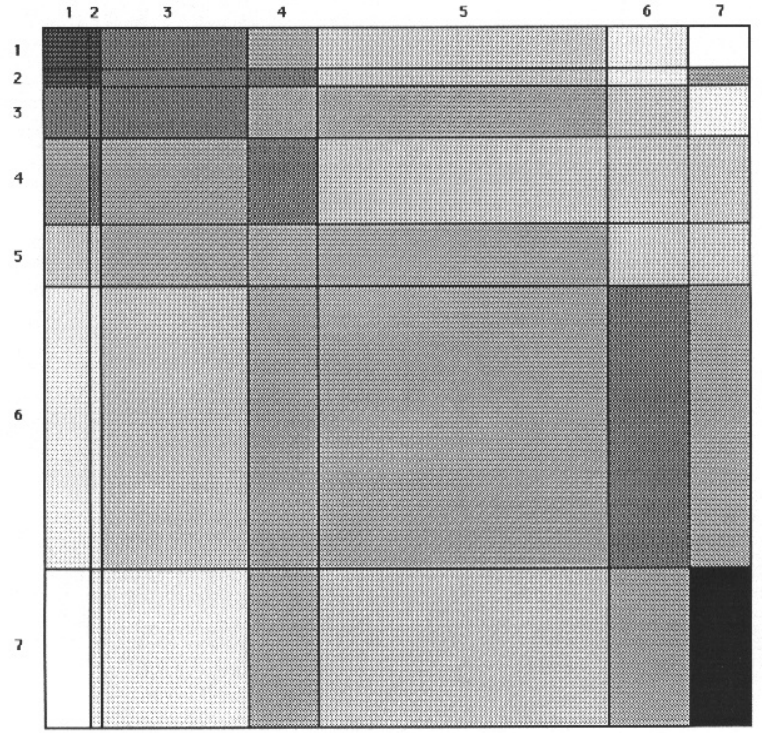
Densité	I	II	III	IV	V	VI	VII
I	1	2,77	6,52	11,68	18,04	30,27	64,80
II	0,36	1	2,35	4,22	6,51	10,93	23,40
III	0,15	0,43	1	1,79	2,77	4,65	9,95
IV	0,09	0,24	0,56	1	1,55	2,59	5,55
V	0,06	0,15	0,36	0,65	1	1,68	3,59
VI	0,03	0,09	0,22	0,39	0,60	1	2,14
VII	0,02	0,04	0,10	0,18	0,28	0,47	1

Lecture : La densité estimée au niveau I vaut 2,77 fois celle au niveau II. De même, la densité estimée au niveau IV vaut 0,56 fois celle au niveau III.

Finalement, ayant obtenu une description statistiquement correcte de la structure de la fluidité sociale, il est possible de la représenter dans la figure suivante…

Modèle final de HAUSER
Représentation graphique des niveaux de densité caractéristiques de la table de mobilité sociale féminine

Source des données : Enquête Formation-Qualification Professionnelle de 1985
Champ : Femmes françaises, actives occupées, âgées de 35 à 59 ans en 1985 (N = 5178)



Les catégories d'origine (lignes du tableau) et de destination (colonnes du tableau) sont repérées dans la nomenclature suivante :

1-Cadres et professions intellectuelles supérieures (sauf Professions libérales)
2-Chefs d'entreprise de 10 salariés ou plus, Professions libérales
3-Professions intermédiaires (sauf Contremaîtres, agents de maîtrise)
4-Artisans, Commerçants et assimilés
5-Employés
6-Contremaîtres, agents de maîtrise, Ouvriers
7-Agriculteurs exploitants

La hauteur de chaque cellule est proportionnelle à l'importance de la catégorie dans la structure sociale des pères. Il en est de même pour la largeur et la structure sociale des filles.

Les différents niveaux de densité sont représentés comme suit :
Valeur du paramètre log-linéaire : 4.17 3.15 2.30 1.71 1.28 0.76 0.00