REVUE DE STATISTIQUE APPLIQUÉE

K. J. WORSLEY

Un exemple d’identification d’un modèle log-linéaire grâce à une analyse des correspondances

Revue de statistique appliquée, tome 35, nº 3 (1987), p. 13-20

<http://www.numdam.org/item?id=RSA_1987__35_3_13_0>

Rev. Statistique Appliquée, 1987, **XXXV** (3), 13-20

UN EXEMPLE D'IDENTIFICATION D'UN MODÈLE LOG-LINÉAIRE GRÂCE À UNE ANALYSE DES CORRESPONDANCES

K.J. WORSLEY

Laboratoire de Statistique et Probabilités, U.A.-C.N.R.S. 745,
Université Paul Sabatier
118, route de Narbonne, 31062 Toulouse Cedex
et Department of Mathematics and Statistics,
McGill University, 805, rue Sherbrooke ouest, Montréal, Québec,
Canada H3A 2K6

RÉSUMÉ

L'analyse des correspondances est souvent proposée comme analyse exploratoire des données, mais les exemples où elle conduit à proposer un modèle log-linéaire sont assez rares. Dans cet article, on présente un tel exemple. Les liens entre les deux approches sont illustrés par un graphique des estimations des paramètres d'un modèle log-linéaire, qui s'avère presque identique à la représentation graphique issue de l'analyse des correspondances.

Mots clés : *Modèles log-linéaires; Analyse des correspondances; GLIM.*

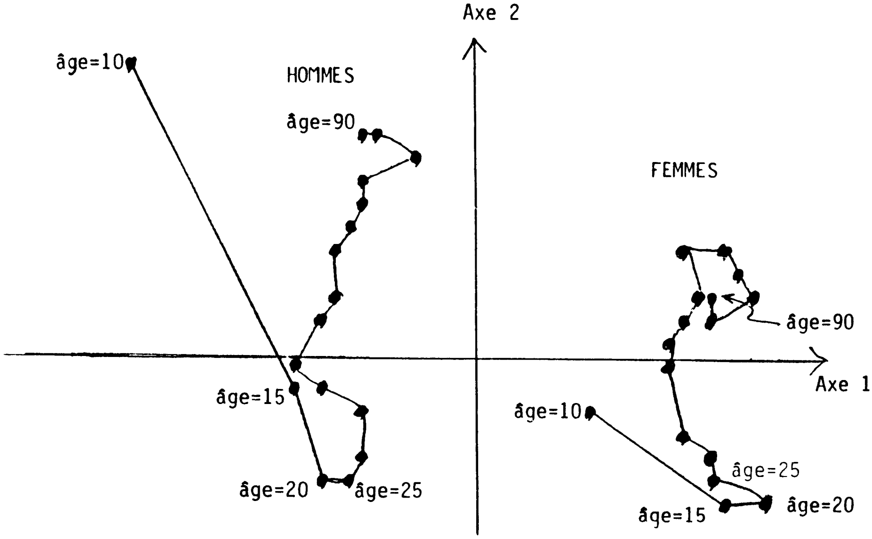
SUMMARY

Correspondence analysis is often proposed as an exploratory data analysis, but examples of where it has lead to the formulation of a log-linear model are rather rare. In this article we shall present such an example. The link between the two approaches is illustrated by a plot of log-linear regression parameter estimates, which turns out to be almost identical to the usual graphical presentation of correspondence analysis scores.

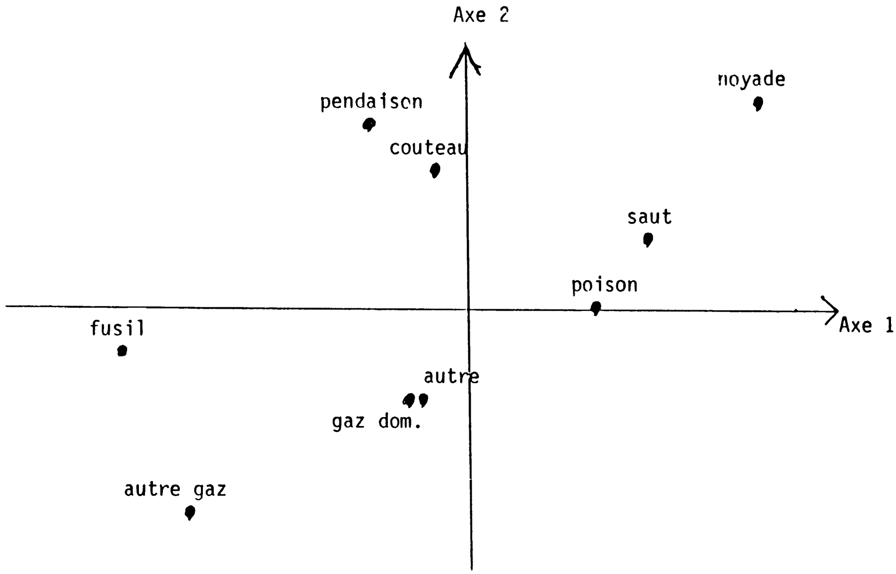
1. Exemple

VAN DER HEIJDEN et DE LEEUW [5] ont développé une analyse des correspondances généralisées pour les résidus d'un modèle log-linéaire. Ils ont illustré leur méthode sur une table de contingence à trois dimensions : le mode de suicide (9 modalités), l'âge des suicidés (17 modalités : 10+, 15+ ,..., 90+) et le sexe (2 modalités). Les données sont les fréquences des suicides constatées en Allemagne durant les années 1974-1977. Dans leur première analyse, les deux variables « âge » et « sexe » ont été combinées en une seule variable « âge-sexe » à 34 modalités, et la table à deux dimensions résultante a été traitée par l'analyse des correspondances simple. Les représentations sur les deux premiers axes sont reproduites sur la figure 1(a) pour les modalités de la variable « âge-sexe », et sur la figure 1(b) pour les modalités de la variable « mode ».

Ces deux axes expliquent 51,9 % et 38,1 % de l'association entre âge-sexe et mode, et les valeurs propres sont respectivement 0,312 et 0,26. Grâce à la



(a) Les modalités de la variable « âge-sexe » :
λ_1 = 0,312 (51,9 %), λ_2 = 0,268 (38,1 %).



(b) Les modalités de la variable « mode ».

FIGURE 1

figure 1(a), on peut constater immédiatement que le premier axe oppose les hommes et les femmes, et que le deuxième axe met en évidence un effet approximativement linéaire de l'âge, excepté les modalités 10+ et 15+ pour les hommes et 10+, 75+ à 90+ pour les femmes. On peut alors se demander s'il y a un modèle log-linéaire susceptible de modéliser ces observations.

2. Liens entre analyse des correspondances et modèles log-linéaires

D'une certaine façon, pour deux dimensions, l'analyse des correspondances simple revient à approcher les fréquences absolues f_{ij} par e_{ij}, avec

$$e_{ij} = cp_i q_j (1 + u_{1i} v_{1j} + u_{2i} v_{2j}) \tag{1}$$

où les paramètres c, p_i, q_j, u_{1i}, v_{1j}, u_{2i}, v_{2j} sont « estimés » par n'importe quelles valeurs minimisant

$$\chi^2 = \sum_{i,j} (f_{ij} - e_{ij})^2 / (f_{i+} f_{+j} / f_{++})$$

où f_{i+}, f_{+j}, f_{++} sont les effectifs marginaux, avec des notations usuelles. Pour plus de clarté dans la suite, la représentation (1) n'est pas écrite sous la forme usuelle. Notons que c, p_i et q_j ne sont pas ici fixés, et qu'il faut imposer des contraintes sur tous les paramètres pour obtenir les coordonnées usuelles de l'analyse des correspondances.

Si u_{1i}, v_{1j}, u_{2i}, v_{2j} sont petits par rapport à 1, on peut écrire

$$\log e_{ij} \approx \mu + \alpha_i + \beta_j + u_{1i} v_{1j} + u_{2i} v_{2j} \tag{2}$$

où $\mu = \log c$, $\alpha_i = \log p_i$ et $\beta_j = \log q_j$, c'est-à-dire un modèle log-multiplicatif pour les fréquences. Cette condition est approximativement satisfaite lorsque les contraintes usuelles sont imposées et que la racine de la somme des carrés des deux premières valeurs propres est petite par rapport à un. C'est le cas dans notre exemple puisque cette valeur est égale à 0,41.

De plus, la minimisation du critère du χ^2 est approximativement équivalente à la minimisation de la déviance

$$D = 2 \sum_{i,j} [f_{ij} \log (f_{ij} / e_{ij}) - (f_{ij} - e_{ij})] ,$$

qui est la méthode du maximum de vraisemblance quand les fréquences sont indépendantes et suivent une loi de Poisson. En ce sens, l'analyse des correspondances est approximativement équivalente à l'ajustement du modèle log-multiplicatif (2) par maximum de vraisemblance (voir ESCOUFIER [2], p. 68, et GOODMAN [4]).

3. Modèles log-linéaires suggérés par l'analyse des correspondances

Pour la table des suicides, soit i l'indice de la modalité âge-sexe et j l'indice de la modalité mode. Les sorties graphiques de l'analyse des correspondances suggèrent que le paramètre u_{1i} prendrait une valeur constante pour les hommes et une autre valeur constante pour les femmes. De plus, le modèle (2) est invariant par transformation linéaire des paramètres u_{1i}. Par suite on peut affecter aux paramètres u_{1i} deux valeurs fixées, disons $x_{1i} = 0$ pour les hommes et $x_{1i} = 1$ pour les femmes. Le terme multiplicatif $u_{1i} v_{1j}$ se transforme alors en un terme linéaire $x_{1i} v_{1j}$, qui est un terme d'interaction entre sexe et mode, au sens du modèle linéaire.

L'analyse des correspondances suggère encore que u_{2i} pourrait prendre des valeurs qui dépendent linéairement de l'âge, mis à part certaines modalités. Comme vu ci-dessus, on peut les remplacer par les valeurs fixées $x_{2i} = 1$ pour l'âge 10+, $x_{2i} = 2$ pour l'âge 15+ ,..., $x_{2i} = 17$ pour l'âge 90+. Le terme $u_{2i} v_{2j}$ se transforme alors en un terme linéaire par rapport à l'âge $x_{2i} v_{2j}$ avec des pentes v_{2j} différentes selon le mode de suicide.

Ainsi, le modèle log-multiplicatif (2) serait approximativement équivalent au modèle log-linéaire

$$\log e_{ij} = \mu + \alpha_i + \beta_j + x_{1i} v_{1j} + x_{2i} v_{2j}. \tag{3}$$

Ce modèle a été ajusté aux fréquences par le logiciel GLIM en trois étapes. On a d'abord ajusté le modèle d'indépendance :

$$\log e_{ij} = \mu + \alpha_i + \beta_j \qquad (D = 10\,330;\ 264 \text{ d.d.l.});$$

puis on a ajouté le terme d'interaction entre sexe et mode et ajusté le modèle :

$$\log e_{ij} = \mu + \alpha_i + \beta_j + x_{1i} v_{1j} \qquad (D = 4\,849;\ 256 \text{ d.d.l.}).$$

La réduction de la déviance est de 53,1 %, soit une valeur très proche de la réduction du χ^2 de 51,9 % associée au premier axe de l'analyse des correspondances.

Enfin les deux premières modalités d'âge ont été supprimées et, en ajoutant le terme $x_{2i} v_{2j}$ linéaire par rapport à l'âge, la déviance du modèle (3) devient $D = 840{,}4$ (216 d.d.l.), soit une réduction de 38,8 %, très proche de la contribution de 38,1 % du deuxième axe de l'analyse des correspondances. On peut justifier la suppression des modalités 10+ et 15+ de l'âge en considérant qu'il est impossible de se fier à leurs fréquences; en effet, mis à part la pendaison, les suicides des jeunes sont souvent attribués à des accidents. De plus, ils correspondent à des fréquences faibles, donc peu significatives.

On peut représenter graphiquement les estimations $(\hat{v}_{1j}, \hat{v}_{2j})$ (voir Figure 2); on constate que ce graphique est presque identique au graphique produit par l'analyse des correspondances, Figure 1(b). Rappelons que l'axe horizontal exprime les effets du sexe et l'axe vertical les effets linéaires de l'âge.

Les estimations elles-mêmes sont données dans la Table I avec leur transformation exponentielle. Notons que $\hat{v}_{11} = 0$ et $\hat{v}_{21} = 0$ à cause des contraintes (arbitraires) imposées par GLIM pour rendre estimables les autres paramètres.

Leur interprétation quantitative est la suivante. Par rapport à l'empoisonnement, les taux des autres modes de suicide sont en général plus faibles chez les femmes; en particulier, le risque relatif de l'utilisation par les femmes des fusils diminue de $100(1 - 0{,}06) = 94$ %. Avec chaque augmentation de l'âge de 5 ans, et toujours par rapport à l'empoisonnement, les modes violents (pendaison, noyade, couteau, saut) sont davantage employés, mis à part les fusils pour lesquels le risque relatif diminue de $100(1 - 0{,}96) = 4$ % chaque 5 ans. Ces conclusions

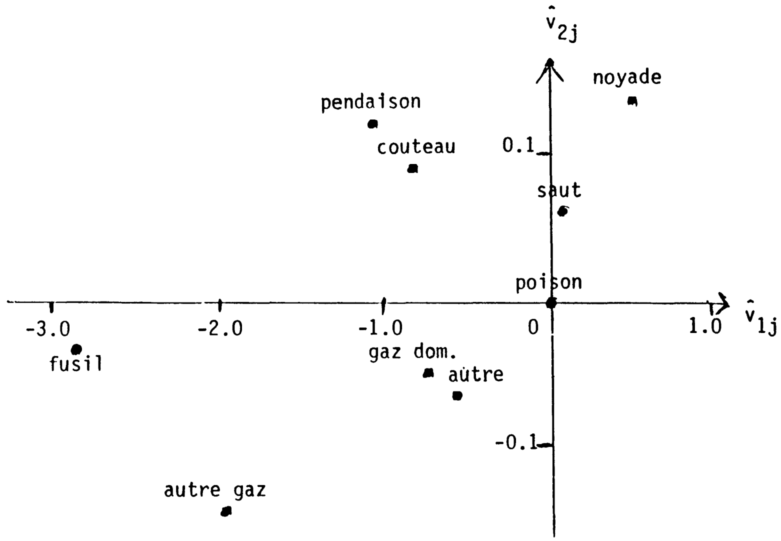


FIGURE 2
Les estimations du modèle log-linéaire (3).

TABLE I

		Effet du sexe (femmes — hommes)		Effet linéaire de l'âge (par 5 ans)	
Mode de suicide	j	$\hat{v}_{1j}$	exp ($\hat{v}_{1j}$)	$\hat{v}_{2j}$	exp ($\hat{v}_{2j}$)
poison	1	0	1	0	1
gaz domestique	2	−0,71	0,49	−0,052	0,95
autre gaz	3	−1,88	0,15	−0,152	0,86
pendaison	4	−1,03	0,36	0,123	1,13
noyade	5	0,47	1,61	0,144	1,15
fusil	6	−2,78	0,06	−0,038	0,96
couteau	7	−0,77	0,46	0,092	1,10
saut	8	0,01	1,10	0,065	1,07
autres	9	−0,56	0,57	−0,068	0,93

quantitatives peuvent servir, par exemple, pour la comparaison avec d'autres pays ou d'autres époques.

L'effet de l'âge est-il vraiment linéaire, c'est-à-dire a-t-on un meilleur ajustement en remplaçant x_{2i} par des paramètres complètement arbitraires, u_{2i} ? Ce problème est étudié en introduisant le modèle log-multiplicatif intermédiaire suivant :

$$\log e_{ij} = \mu + \alpha_i + \beta_j + x_{1i} v_{1j} + u_{2i} v_{2j} . \tag{4}$$

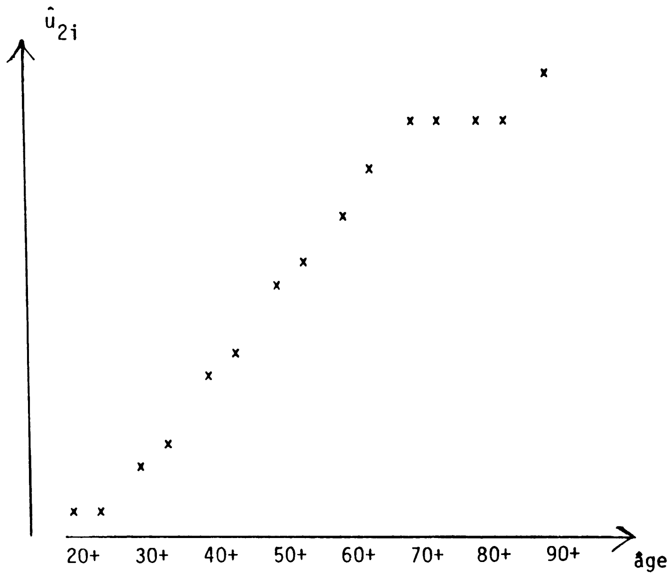


FIGURE 3
Les estimations du modèle log-multiplicatif intermédiaire (4).

Ce modèle a été ajusté par la méthode itérative proposée par BREEN [1] : en fixant les valeurs des u_{2i} on estime les v_{2j}, puis en donnant aux v_{2j} ces valeurs on estime à leur tour les u_{2i}, et on itère. A chaque itération on ajuste un modèle log-linéaire, et trois directives de GLIM suffisent pour cela. Après quatre itérations la déviance est stabilisée à la valeur de 765,4 (203 d.d.l.). La Figure 3 donne une représentation graphique des estimations $\hat{u}_{2i}$ en fonction de l'âge. On peut constater que c'est à peu près linéaire, sauf peut-être pour les âges 70+ à 90+ qui sont plus proches que les autres; cette proximité est aussi présente dans la Figure 1(a). Cette constatation correspond à la faible diminution de la déviance par rapport à la déviance du modèle (3), linéaire par rapport à l'âge. On pourrait donc attribuer une grande partie de la différence entre ces deux modèles (3) et (4) aux âges 70+ à 90+.

La méthode itérative pour ajuster les modèles multiplicatifs semble rapide et efficace. En effet l'analyse des correspondances de départ et les graphiques des Figures 1(a) et 1(b) ont été obtenus par GLIM en utilisant la même méthode itérative, mais en remplaçant le critère du maximum de vraisemblance, D, par le critère des moindres carrés pondérés, χ^2. Dans ce cas, cet algorithme est équivalent à l'algorithme des « moyenne réciproques » [4].

4. Approche modélisatrice classique

L'analyse des correspondances nous a conduit directement à un modèle log-linéaire qui s'ajuste bien aux données avec moins de paramètres qu'un modèle log-multiplicatif. On peut se demander si on aurait été guidé vers un tel modèle par les méthodes classiques. Devant cette table à trois dimensions, il est naturel de considérer le modèle avec toutes les interactions entre deux facteurs (D = 429; 128 d.d.l.). Mais le facteur âge étant ordonné, on le remplace souvent dans les interactions par une variable linéaire. On retrouverait alors le modèle (3), les modalités 10+ et 15+ de l'âge comprises; ce modèle a une déviance de 1 741 (248 d.d.l.). Une dernière question se pose : est-ce que l'on peut mettre en évidence que les deux modalités 10+ et 15+ sont aberrantes ? Un graphique donnant les résidus standardisés en fonction de l'âge (Figure 4) permet de constater que les résidus associés à ces modalités sont anormalement grands, et on pourrait donc les supprimer.

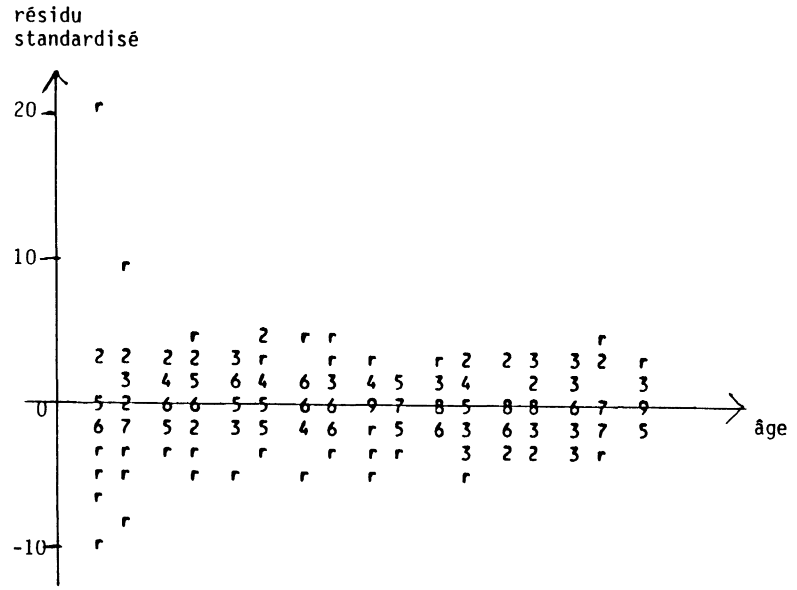


FIGURE 4

Les résidus standardisés du modèle (3), les modalités 10+ et 15+ de l'âge comprise.

On peut, bien sûr, chercher des modèles plus raffinés pour réduire la déviance. Par exemple, en ajoutant un terme d'interaction entre le mode de suicide et une variable quadratique par rapport à l'âge $x_{2i}^2 v_{3j}$, la déviance devient D = 483,2 (208 d.d.l.). En principe, on peut continuer de cette manière jusqu'à ce que la déviance soit acceptée par un test de χ^2 au niveau, disons, de 5 %. Pourtant, tous les modèles déjà considérés ne sont pas acceptables; il y a peut-être trois raisons pour cela. D'abord, les fréquences pourraient dépendre d'autres variables non mesurées (habitation urbaine ou rurale, proximité d'une surface d'eau, saison de l'année...); il est en effet difficile d'admettre que les seules variables âge, sexe et mode peuvent complètement expliquer les distributions des fréquences. Deuxièmement, si ces variables non mesurées sont aléatoires, un modèle utilisant les variables âge, sexe et mode pourrait être satisfaisant pour les moyennes des fréquences, mais en revanche les fréquences ne suivront plus une loi de Poisson; en particulier les variances des fréquences augmenteront, ainsi que les valeurs critiques de la loi nulle de la déviance. Enfin, des corrélations entre les fréquences produiront le même effet; il est bien connu qu'un suicide peut en provoquer un autre.

5. Conclusion

Un statisticien adroit aurait donc pu retrouver le modèle (3) par les méthodes classiques; mais, en revanche, l'analyse des correspondances nous dirige directement, et avec plus de sûreté, vers un modèle log-linéaire approprié, un modèle qui peut fournir des conclusions quantitatives complétant utilement les représentations graphiques de l'analyse des correspondances.

Bibliographie

[1] R. BREEN (1984). — Fitting non-hierarchical association log-linear model using GLIM. *Sociological Methods and Research*, Vol. 13, 77-107.

[2] Y. ESCOUFIER (1982). — L'analyse des tableaux de contingence simples et multiples. *Metron*, Vol. XL, 53-77.

[3] L.A. GOODMAN (1985). — The analysis of cross-classified data having ordered and/or unordered catagories : Association models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, Vol. 13, 10-69.

[4] M.O. HILL (1974). — Correspondence analysis : a neglected multivariate method. *Applied Statistics*, Vol. 23, 340-354.

[5] P.G.M. van der HEIJDEN et J. de LEEUW (1985). — Correspondence anlaysis used complementary to log-linear analysis. *Psychometrika*, Vol. 50, 429-448.