

**L'Estimation de Modèles Log-Linéaires
sur des Tableaux de Contingence issus
d'Enquêtes à plan de sondage complexe :**

**un Examen de l'Approche proposée par
Clogg & Eliason**

Chris Skinner
Université de Southampton

Louis-André Vallet
CNRS & CREST, Paris

Plan

- modèles log-linéaires et échantillonnage
- l'approche de
Clogg C. C. and Eliason S. R. (1987) Some common problems in log-linear analysis, *Sociological Methods & Research*, 16, 8-44.
- approche de pseudo-maximum de vraisemblance
- comparaison théorique

- étude empirique de mobilité sociale sur un tableau issu de l'enquête *Formation et Qualification Professionnelle* (1985)

Tableau de Contingence (au niveau de la population)

| | | |
|----------|----------|----------|
| N_{11} | N_{12} | N_{13} |
| N_{21} | N_{22} | N_{23} |
| N_{31} | N_{32} | N_{33} |

Modèle Log-Linéaire

| | | |
|------------|------------|------------|
| μ_{11} | μ_{12} | μ_{13} |
| μ_{21} | μ_{22} | μ_{23} |
| μ_{31} | μ_{32} | μ_{33} |

$$E(N_{ij}) = \mu_{ij}, \quad \log(\mu_{ij}) = \alpha + \beta_i + \gamma_j$$

Échantillonnage

Deux cas :

- A. probabilités de sélection constantes ($= \pi_{ij}$) à l'intérieur des cellules ij
- B. probabilités de sélection variables

n_{ij} taille de l'échantillon dans la cellule ij

\hat{N}_{ij} somme des poids d'échantillonnage dans cellule ij

Approche de Clogg & Eliason

Cas A : poids constants à l'intérieur des cellules

- $E(n_{ij}) = \mu_{sij} = \pi_{ij} E(N_{ij}) = \pi_{ij} \mu_{ij}$
- $\log(\mu_{sij}) = \log(\pi_{ij}) + \alpha + \beta_i + \gamma_j$
- $\log(\pi_{ij})$ 'offset' du modèle
- estimation par maximum de vraisemblance

Approche de Clogg & Eliason

Cas B: poids variables (cas général)

- π_{ij} estimé par $\hat{\pi}_{ij} = (\text{poids moyen})^{-1}$
- $\log(\hat{\pi}_{ij})$ 'offset' du modèle
- estimation par maximum de vraisemblance

Approche du Pseudo-Maximum de Vraisemblance (PMV)

- estimation ponctuelle avec \hat{N}_{ij}
- estimation des erreurs-types par méthodes de sondage (linéarisation, jackknife, etc.)

Comparaison Théorique de CE & PMV

- estimateurs ponctuels CE & PMV différents
- chaque estimateur ponctuel est sans biais (approx.) si modèle vrai
- estimation CE des erreurs-types non valide, sauf en cas A ; sous-estimation systématique en général ; ignore l'échantillonnage en grappes
- estimation PMV des erreurs-types valide

Comparaison Empirique

- Enquête Insee *Formation & Qualification Professionnelle* 1985
- Population (approx.) : personnes des ménages ordinaires âgées de 13 à 69 ans au recensement de 1982
- Échantillon : 46 500 personnes par échantillonnage stratifié (73 strates) avec des fractions de sondage comprises entre $1/2690$ et $1/200$
- 39 233 répondants
- La variable de poids reflète à la fois les probabilités inégales d'inclusion et la non-réponse à l'enquête

Données de mobilité sociale analysées

Sous-échantillon de 5 159 femmes, âgées de 35 à 59 ans en 1985, actives occupées à la date d'enquête,

pour lesquelles on connaît :

- la catégorie socioprofessionnelle ;
- et celle de leur père (quand elles ont cessé de fréquenter régulièrement l'école ou l'université)

(distribution de l'échantillon analysé dans les différentes strates)

| Strate (situation en 1982) | n | fraction de sondage | poids moyen | écart-type du poids |
|---|--------------|---------------------|-------------|---------------------|
| French, in labour market, farmers, 32-51 | 234 | 1/940 | 960 | 124 |
| French, in labour market, farmers, 52+ | 83 | 1/1250 | 1246 | 48 |
| French, in labour market, artisans/shopkeepers, 32-51 | 223 | 1/1040 | 1145 | 88 |
| French, in labour market, artisans/shopkeepers, 52+ | 28 | 1/1360 | 1488 | 120 |
| French, in labour market, managers/high professional, 32-51 | 747 | 1/310 | 344 | 44 |
| French, in labour market, managers/high professionals, 52+ | 94 | 1/340 | 389 | 67 |
| French, in labour market, low professionals, 32-51 | 1 064 | 1/600 | 669 | 112 |
| French, in labour market, low professionals, 52+ | 101 | 1/620 | 720 | 76 |
| French, in labour market, non manual, 32-51 | 1 581 | 1/830 | 935 | 129 |
| French, in labour market, non manual, 52+ | 214 | 1/830 | 946 | 112 |
| French, in labour market, manual, 32 to 51 | 535 | 1/760 | 840 | 75 |
| French, in labour market, manual, 52+ | 60 | 1/1080 | 1194 | 51 |
| French, in labour market, unemployed & never worked | 13 | 1/400 | 492 | 93 |
| French, students | 7 | 1/900 | 1000 | 111 |
| French, previously in the labour market | 2 | 1/2270 | 2464 | 247 |
| Other French, out of labour market, 32-51 | 146 | 1/2500 | 2795 | 621 |
| Other French, out of labour market, 52+ | 17 | 1/2500 | 2795 | 389 |
| Foreign, in labour market, employed or unemployed, 32-51 | 10 | 1/730 | 831 | 190 |
| Total | 5 159 | - | 850 | 451 |

(table de mobilité)

| Daughter's class | Freq. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total |
|--|--------------|------------------|-----------------|--------------------|------------------|--------------------|------------------|------------------|--------------------|
| Father's class | | | | | | | | | |
| 1 Higher-grade salaried professionals | Unw. Wei. | 164.00 81.23 | 25.00 13.01 | 136.00 113.18 | 12.00 15.35 | 59.00 66.32 | 9.00 8.08 | 0.00 0.00 | 405.00 297.17 |
| 2 Company managers and liberal professions | Unw. Wei. | 56.00 28.78 | 27.00 11.72 | 37.00 38.22 | 14.00 14.46 | 28.00 32.45 | 3.00 2.65 | 3.00 7.01 | 168.00 135.29 |
| 3 Lower-grade salaried professionals | Unw. Wei. | 95.00 48.08 | 16.00 11.44 | 161.00 129.70 | 15.00 22.79 | 115.00 131.79 | 18.00 18.20 | 4.00 4.77 | 424.00 366.78 |
| 4 Artisans and shopkeepers | Unw. Wei. | 97.00 52.25 | 35.00 21.35 | 219.00 174.45 | 78.00 118.41 | 200.00 223.37 | 35.00 39.57 | 8.00 14.27 | 672.00 643.67 |
| 5 Non-manual workers | Unw. Wei. | 59.00 30.18 | 7.00 3.68 | 145.00 120.03 | 32.00 53.42 | 182.00 216.57 | 29.00 28.65 | 3.00 4.17 | 457.00 456.70 |
| 6 Foremen and manual workers | Unw. Wei. | 128.00 64.18 | 18.00 14.88 | 419.00 361.46 | 124.00 184.12 | 930.00 1065.19 | 339.00 355.76 | 37.00 47.06 | 1995.00 2092.66 |
| 7 Farmers | Unw. Wei. | 38.00 20.29 | 8.00 5.63 | 164.00 134.71 | 73.00 101.98 | 342.00 394.83 | 136.00 140.49 | 277.00 368.80 | 1038.00 1166.73 |
| Total | Unw. Wei. | 637.00 324.99 | 136.00 81.71 | 1281.00 1071.75 | 348.00 510.54 | 1856.00 2130.52 | 569.00 593.40 | 332.00 446.08 | 5159.00 5159.00 |

Analyser la structure et la force de l'association : le modèle log-linéaire de Hauser (1978)

Il identifie les effets d'association entre les deux variables en contraignant certains d'entre eux à être égaux pour des ensembles de cellules du tableau de contingence.

On suppose que :

les cellules ij sont assignées à K sous-ensembles ;

et chacun partage un même paramètre d'association δ_k .

D'où le modèle :
$$\text{Log } m_{ij} = \alpha + \beta_i + \gamma_j + \delta_k$$
 si la cellule ij appartient au sous-ensemble k

Les paramètres δ_k reflètent la densité de mobilité ou d'immobilité dans certaines cellules (relativement à d'autres cellules du tableau).

Modèle initial et modèle final du tableau de mobilité

Dans un travail précédent (JMS 2005), sur la base d'hypothèses sociologiques,

nous avons proposé un tel modèle (ou allocation des cellules) fondé sur $K=7$ paramètres d'association.

Il s'est avéré relativement proche des données observées.

Après quelques modifications, il en résulte un modèle final (avec, de nouveau, $K=7$ paramètres d'association) qui s'ajuste de façon satisfaisante aux données (au sens d'un test statistique).

| | <i>Modèle initial</i> | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|--|-----------------------|-----|-----|-----|----|----|-----|-----|
| 1 – Higher-grade salaried professionals | | II | III | IV | V | VI | VII | VII |
| 2 – Company managers and liberal professions | | III | II | IV | IV | VI | VII | VII |
| 3 – Lower-grade salaried professionals | | IV | IV | IV | V | V | VI | VII |
| 4 – Artisans and shopkeepers | | V | IV | V | IV | V | VI | VI |
| 5 – Non-manual workers | | VI | VI | V | V | V | V | VI |
| 6 – Foremen and manual workers | | VII | VII | VI | VI | V | IV | V |
| 7 – Farmers | | VII | VII | VII | VI | VI | V | I |

Parmi les effets d'association, I est supposé être le plus fort et VII le plus faible.

Comparaison des estimateurs ponctuels et des erreurs-types dans quatre analyses

- Quand les effectifs non pondérés sont analysés
- Quand les effectifs pondérés sont analysés en ignorant la complexité du plan de sondage
- Avec l'approche proposée par Clogg & Eliason

ces trois analyses avec les procédures SAS Catmod et Genmod

- Avec l'approche du pseudo-maximum de vraisemblance avec le logiciel IVEware et la procédure SAS Catmod méthode de "Jackknife Repeated Replication" (JRR) pour les estimations de variance
intensif du point de vue du calcul : environ 50 minutes pour chaque modèle

| Paramètre | Modèle initial | | | | Modèle final | | | |
|-----------------|-------------------|-------------------|-------------------|---------------------------|-------------------|-------------------|-------------------|---------------------------|
| | Non pondéré | Pondéré | Clogg & Eliason | Pseudo maximum likelihood | Non pondéré | Pondéré | Clogg & Eliason | Pseudo maximum likelihood |
| β_1 (se) | -1.813 (0.087) | -1.825 (0.086) | -1.828 (0.086) | -1.825 (0.098) | -1.747 (0.084) | -1.754 (0.083) | -1.763 (0.083) | -1.754 (0.093) |
| β_2 (se) | -2.626 (0.107) | -2.621 (0.108) | -2.612 (0.106) | -2.621 (0.133) | -2.663 (0.102) | -2.610 (0.105) | -2.632 (0.102) | -2.610 (0.125) |
| β_3 (se) | -1.532 (0.079) | -1.559 (0.078) | -1.549 (0.079) | -1.559 (0.090) | -1.492 (0.076) | -1.517 (0.075) | -1.514 (0.076) | -1.517 (0.085) |
| β_4 (se) | -0.856 (0.069) | -0.857 (0.067) | -0.855 (0.070) | -0.857 (0.079) | -0.633 (0.061) | -0.614 (0.059) | -0.643 (0.061) | -0.614 (0.068) |
| β_5 (se) | -1.134 (0.072) | -1.104 (0.072) | -1.111 (0.073) | -1.104 (0.082) | -1.036 (0.067) | -1.013 (0.065) | -1.021 (0.067) | -1.013 (0.075) |
| β_6 (se) | 0.492 (0.049) | 0.510 (0.049) | 0.505 (0.049) | 0.510 (0.056) | 0.487 (0.048) | 0.507 (0.047) | 0.497 (0.048) | 0.507 (0.054) |
| β_7 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 |
| γ_1 (se) | 2.187 (0.149) | 1.179 (0.139) | 1.261 (0.149) | 1.179 (0.166) | 2.177 (0.148) | 1.196 (0.138) | 1.238 (0.148) | 1.196 (0.157) |
| γ_2 (se) | 0.585 (0.169) | -0.269 (0.169) | -0.182 (0.170) | -0.269 (0.205) | 0.450 (0.167) | -0.373 (0.167) | -0.321 (0.167) | -0.373 (0.198) |
| γ_3 (se) | 2.889 (0.140) | 2.360 (0.120) | 2.424 (0.140) | 2.360 (0.150) | 2.855 (0.139) | 2.341 (0.119) | 2.376 (0.139) | 2.341 (0.146) |

| | | | | | | | | |
|---------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| γ_4 (se) | 1.473 (0.147) | 1.508 (0.124) | 1.555 (0.148) | 1.508 (0.156) | 1.204 (0.147) | 1.253 (0.124) | 1.282 (0.147) | 1.253 (0.153) |
| γ_5 (se) | 3.089 (0.137) | 2.895 (0.116) | 2.943 (0.137) | 2.895 (0.143) | 3.167 (0.137) | 2.971 (0.116) | 3.003 (0.137) | 2.971 (0.144) |
| γ_6 (se) | 1.605 (0.146) | 1.297 (0.126) | 1.349 (0.146) | 1.297 (0.150) | 1.638 (0.146) | 1.340 (0.126) | 1.370 (0.146) | 1.340 (0.150) |
| γ_7 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 |
| δ_I (se) | 3.561 (0.163) | 3.451 (0.146) | 3.569 (0.163) | 3.451 (0.189) | 4.163 (0.228) | 4.096 (0.266) | 4.138 (0.228) | 4.096 (0.252) |
| δ_{II} (se) | 2.730 (0.119) | 2.619 (0.147) | 2.660 (0.118) | 2.619 (0.135) | 3.215 (0.191) | 3.104 (0.251) | 3.123 (0.191) | 3.104 (0.214) |
| δ_{III} (se) | 2.396 (0.150) | 2.297 (0.189) | 2.326 (0.149) | 2.297 (0.186) | 2.276 (0.187) | 2.252 (0.245) | 2.275 (0.187) | 2.252 (0.208) |
| δ_{IV} (se) | 1.683 (0.086) | 1.633 (0.093) | 1.700 (0.085) | 1.633 (0.105) | 1.692 (0.183) | 1.658 (0.243) | 1.675 (0.183) | 1.658 (0.204) |
| δ_V (se) | 1.161 (0.084) | 1.078 (0.092) | 1.154 (0.084) | 1.078 (0.103) | 1.245 (0.181) | 1.217 (0.241) | 1.240 (0.181) | 1.217 (0.201) |
| δ_{VI} (se) | 0.683 (0.072) | 0.641 (0.080) | 0.699 (0.072) | 0.641 (0.087) | 0.731 (0.177) | 0.708 (0.239) | 0.702 (0.177) | 0.708 (0.196) |
| δ_{VII} | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 | Fixed at 0 |
| Déviante | 86.11 | 77.12 | 75.58 | - | 47.71 | 33.69 | 34.77 | - |
| DDL | 29 | 29 | 29 | - | 29 | 29 | 29 | - |

Estimateurs ponctuels : résultats

- Ceux obtenus en analysant les effectifs non pondérés sont biaisés et peuvent être nettement différents de tous les autres.

(en particulier les paramètres relatifs à la variable-colonne (CS de la fille en 1985) car elle est fortement liée à l'une des variables de stratification)

- Ceux obtenus en analysant les effectifs pondérés et sous l'approche du pseudo-maximum de vraisemblance sont identiques comme attendu.
- Ceux obtenus avec l'approche de Clogg & Eliason sont proches des précédents, bien que non exactement semblables.

Erreurs-types : résultats

- Comme attendu, elles sont différentes entre l'approche PMV (qui prend en compte le plan de sondage) et l'approche pondérée (qui ne le fait pas).
- Sous l'approche Clogg & Eliason, elles sont virtuellement identiques à celle de l'approche non pondérée, mais inférieures à celles de l'approche PMV.
- Or, ces dernières sont très proches de ce que fournit une estimation correcte (jackknife) des erreurs-types sous l'approche CE.
- On retrouve donc empiriquement ce qu'indiquait la comparaison sur le plan théorique : les erreurs-types obtenues sous l'approche Clogg & Eliason sous-estiment généralement la variabilité vraie des paramètres.