Modèle log-linéaire ou poissonien

Résumé

Introductions au modèle linéaire et modèle linéaire général. Le modèle poissonien ou log-linéaire.

Retour au plan du cours.

1 Introduction

Comme dans le chapitre précédent, les modèles décrits dans ce chapitre s'intéressent plus particulièrement à la description ou l'explication d'observations constitués d'effectifs ; nombre de succès d'une variable de Bernouilli lors d'une séquence d'essais dans la cas précédent de la régression logistique, nombre d'individus qui prennent une combinaison donnée de modalités de variables qualitatives ou niveaux de facteurs, dans le cas présent. Ce modèle fait également partie de la famille du *modèle linéaire général* en étant associé à une loi de Poisson. Il est également appelé aussi *modèle log-linéaire* (voir Agresti (1990) pour un exposé détaillé) et s'applique principalement à la modélisation d'une table de contingence complète. Comme pour la régression logistique, les aspects au modèle linéaire général (estimation, tests, diagnostic) ont des stratégies de mise en œuvreest similaire au cas gaussien ; ils ne sont pas repris.

2 Modèle log-linéaire

2.1 Types de données

Les données se présentent généralement sous la forme d'une table de contingence obtenue par le croisement de plusieurs variables qualitatives et dont chaque cellule contient un effectif ou une fréquence à modéliser. Nous nous limiterons à l'étude d'une table élémentaire en laissant de côté des structures plus complexes, par exemple lorsque des zéros structurels, des indépendances conditionnelles, des propriétés de symétrie ou quasi-symétrie, une table creuse, sont à prendre en compte. D'autre part, sous sa forme la plus générale, le modèle peut intégrer également des variables quantitatives.

Ce type de situation se retrouve en analyse des correspondances simple ou multiple mais ici, l'objectif est d'expliquer ou de modéliser les effectifs en fonction des modalités prises par les variables qualitatives. L'objectif final pouvant être *explicatif* : tester une structure de dépendance particulière, ou *prédictif* avec choix d'un modèle parcimonieux.

2.2 Distributions

On considère la table de contingence complète constituée à partir de l'observation des variables qualitatives $X^1, X^2, \ldots, X^p$ sur un échantillon de n individus. Les effectifs $\{y_{jk\ldots l}; j = 1, J; k = 1, K; \ldots; l = 1, L\}$ de chaque cellule sont rangés dans un vecteur $\mathbf{y}$ à $I(I = J \times K \times \cdots \times L)$ composantes. Différentes hypothèses sur les distributions sont considérées en fonction du contexte expérimental.

Poisson

Le modèle le plus simple consiste à supposer que les variables observées Y_i suivent des lois de Poisson indépendantes de paramètre $\mu_i = E(Y_i)$. La distribution conjointe admet alors pour densité :

$$f(\mathbf{y}, \mu) = \prod_{i=1}^{I} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}.$$

La somme $N(N = y_+ = \sum_i y_i)$ des I variables aléatoires de Poisson indépendantes est également une variable de Poisson de paramètre $\mu_+ = \sum_i \mu_i$.

Multinomiale

En pratique, le nombre total n d'observations est souvent fixé a priori par l'expérimentateur et ceci induit une contrainte sur la somme des y_i. La distribution conjointe des variables Y_i est alors conditionnée par n et la densité devient :

$$f(\mathbf{y}, \mu) = \prod_{i=1}^{I} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \Big/ \frac{\mu_+^n e^{-\mu_+}}{n!}.$$

Comme $\mu_+^n = \sum_i \mu_+^{y_i}$ et $e^{-\mu_+} = \prod_i e^{-\mu_i}$, en posant $\pi_i = \frac{\mu_i}{\mu_+}$, on obtient :

$$f(\mathbf{y}, \mu) = n! \prod_{i=1}^{I} \frac{\pi_i^{y_i}}{y_i!} \quad \text{avec} \quad \sum_{i=1}^{I} \pi_i = 1 \text{ et } 0 \leq \pi_i \leq 1; i = 1, I.$$

On vérifie donc que $f(\mathbf{y}, \mu)$ est la fonction de densité d'une loi multinomiale dans laquelle les paramètres π_i modélisent les probabilités d'occurrence associées à chaque cellule. Dans ce cas, $E(Y_i) = n\pi_i$.

Produit de multinomiales

Dans d'autres circonstances, des effectifs marginaux lignes, colonnes ou sous-tables, peuvent être également fixés par l'expérimentateur comme dans le cas d'un sondage stratifié. Cela correspond au cas où une ou plusieurs variables sont contrôlées et ont donc un rôle explicatif ; leurs modalités sont connues *a priori*. Les lois de chacun des sous-éléments de la table, conditionnées par l'effectif marginal correspondant sont multinomiales. La loi conjointe de l'ensemble est alors un produit de multinomiales.

Conséquence

Trois modèles de distribution : Poisson, multinomial, produit de multinomiales, sont envisageables pour modéliser Y_i en fonction des conditions expérimentales. D'un point de vue théorique, on montre que ces modèles conduisent aux mêmes estimations des paramètres par maximum de vraisemblance. La différence introduite par le conditionnement intervient par une contrainte qui impose la présence de certains paramètres dans le modèle, ceux reconstruisant les marges fixées.

2.3 Modèles à 2 variables

Soit une table de contingence $(J \times K)$ issue du croisement de deux variables qualitatives X^1 à J modalités et X^2 à K modalités et dont l'effectif total n est fixé. La loi conjointe des effectifs Y_{jk} de chaque cellule est une loi multinomiale de paramètre π_{jk} et d'espérance :

$$E(Y_{jk}) = n\pi_{jk}.$$

Par définition, les variables X^1 et X^2 sont *indépendantes* si et seulement si :

$$\pi_{jk} = \pi_{+k}\pi_{j+}$$

où π_{j+} (resp. π_{+k}) désigne la loi marginale de X^1 (resp. X^2) :

$$\pi_{j+} = \sum_{k=1}^{K} \pi_{jk} \quad \text{et} \quad \pi_{+k} = \sum_{j=1}^{J} \pi_{jk}.$$

Si l'indépendance n'est pas vérifiée, on peut décomposer :

$$E(Y_{jk}) = n\pi_{jk} = n\pi_{j+}\pi_{+k}\frac{\pi_{jk}}{\pi_{j+}\pi_{+k}}.$$

Notons $\eta_{jk} = \ln(E(Y_{jk}))$. L'intervention de la fonction logarithme permet de linéariser la décomposition précédente autour du "modèle d'indépendance" :

$$\eta_{jk} = \ln n + \ln \pi_{j+} + \ln \pi_{+k} + \ln\left(\frac{\pi_{jk}}{\pi_{j+}\pi_{+k}}\right).$$

Ce modèle est dit *saturé* car, présentant autant de paramètres que de données, il explique exactement celles-ci. L'indépendance est vérifiée si le dernier terme de cette expression, exprimant une dépendance ou interaction comme dans le modèle d'analyse de variance, est nul pour tout couple (j, k).

Les logiciels mettent en place d'autres paramétrisations en faisant apparaître des effets différentiels, soit par rapport à une moyenne, soit par rapport à la dernière modalité.

Dans le premier cas, en posant :

$$\begin{aligned}
\beta_0 &= \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\eta_{jk} = \eta_{..}, \\
\beta_j^1 &= \frac{1}{K}\sum_{k=1}^{K}\eta_{jk} - \eta_{..} = \eta_{j.} - \eta_{..}, \\
\beta_k^2 &= \frac{1}{J}\sum_{j=1}^{J}\eta_{jk} - \eta_{..} = \eta_{.k} - \eta_{..}, \\
\beta_{jk}^{12} &= \eta_{jk} - \eta_{j.} - \eta_{.k} + \eta_{..},
\end{aligned}$$

avec les relations :

$$\forall j, \forall k, \sum_{j=1}^{J} \beta_j^1 = \sum_{k=1}^{K} \beta_k^2 = \sum_{j=1}^{J} \beta_{jk}^{12} = \sum_{k=1}^{K} \beta_{jk}^{12} = 0,$$

le modèle saturé s'écrit :

$$\ln(E(Y_{jk})) = \eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_{jk}^{12}.$$

Il se met sous la forme matricielle

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

où $\mathbf{X}$ est la matrice expérimentale (design matrix) contenant les indicatrices. L'indépendance est obtenue lorsque tous les termes d'interaction β_{jk}^{12} sont nuls.

La deuxième paramétrisation considère la décomposition :

$$\pi_{jk} = \pi_{JK} \frac{\pi_{Jk}}{\pi_{JK}} \frac{\pi_{jK}}{\pi_{JK}} \frac{\pi_{jk}\pi_{JK}}{\pi_{Jk}\pi_{jK}}.$$

En posant :

$$\begin{aligned}
\beta_0 &= \ln n + \ln \pi_{JK}, \\
\beta_j^1 &= \ln \pi_{jK} - \ln \pi_{JK}, \\
\beta_k^2 &= \ln \pi_{Jk} - \ln \pi_{JK}, \\
\beta_{jk}^{12} &= \ln \pi_{jk} - \ln \pi_{jK} - \ln \pi_{Jk} + \ln \pi_{JK},
\end{aligned}$$

avec les mêmes relations entre les paramètres. Le modèle se met encore sous la forme :

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

et se ramène à l'indépendance si tous les paramètres β_{jk}^{12} sont nuls.

Si l'hypothèse d'indépendance est vérifiée, on peut encore analyser les effets principaux :

$$\text{si}, \forall j, \beta_j^1 = 0 \quad \text{alors}, \quad \pi_{jk} = \pi_{Jk} = \frac{1}{J}\pi_{+k}.$$

Il y a équiprobabilité des modalités de X^1. Même chose avec X^2 si les termes β_k^2 sont tous nuls.

Les paramètres du modèle log-linéaire sont estimés en maximisant la log-vraisemblance dont l'explicitation est reportée au chapitre suivant comme cas particulier de modèle linéaire généralisé. Pour les modèles simples, les estimations sont déduites des effectifs marginaux mais comme, dès que le modèle est plus compliqué, des méthodes itératives sont nécessaires, elles sont systématiquement mises en œuvre.

2.4 Modèle à trois variables

On considère une table de contingence $(J \times K \times L)$ obtenue par croisement de trois variables qualitatives X^1, X^2, X^3. La définition des paramètres est conduite de manière analogue au cas de deux variables en faisant apparaître des effets principaux et des interactions. Le modèle saturé se met sous la forme :

$$\ln(E(Y_{jkl})) = \eta_{jkl} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13} + \beta_{kl}^{23} + \beta_{jkl}^{123}$$

et peut aussi est présenté sous forme matricielle.

Nous allons expliciter les sous-modèles obtenus par nullité de certains paramètres et qui correspondent à des structures particulières d'indépendance. Une façon classique de nommer les modèles consiste à ne citer que les interactions retenues les plus complexes. Les autres, ainsi que les effets principaux, sont contenues de par la structure hiérarchique du modèle. Ainsi, le modèle saturé est désigné par $(X^1X^2X^3)$ correspondant à la syntaxe `X1|X2|X3` de SAS.

2.4.1 Cas poissonnien ou multinomial

Seul le nombre total d'observations n est fixé dans le cas multinomial, ceci impose simplement la présence de β_0 dans le modèle.

1. Modèle partiel d'association ou de tout interaction d'ordre 2 : (X^1X^2, X^2X^3, X^1X^3)

   Les termes β_{jkl}^{123} sont tous nuls, seules les interactions d'ordre 2 sont présentes. C'est le modèle implicitement considéré par l'analyse multiple des correspondances. Il s'écrit :

   $$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13} + \beta_{kl}^{23}.$$

2. Indépendance conditionnelle : (X^1X^2, X^1X^3)

Si, en plus, l'un des termes d'interaction est nul, par exemple $\beta_{kl} = 0$ pour tout couple (k, l), on dit que X^2 et X^3 sont indépendantes conditionnellement à X^1 et le modèle devient :

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{jk}^{12} + \beta_{jl}^{13}.$$

3. Variable indépendante : (X^1, X^2X^3)
   Si deux termes d'interaction sont nuls : $\beta_{jl}\beta_{jk} = 0$ pour tout triplet (j, k, l), alors X^1 est indépendante de X^2 et X^3.

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3 + \beta_{kl}^{23}.$$

4. Indépendance : (X^1, X^2, X^3)
   Tous les termes d'interaction sont nuls :

$$\eta_{jk} = \beta_0 + \beta_j^1 + \beta_k^2 + \beta_l^3$$

   et les variables sont mutuellement indépendantes.

2.4.2 *Produit de multinomiales*

– Si une variable est explicative, par exemple X^3, ses marges sont fixées, le modèle doit nécessairement conserver les paramètres

$$\eta_{jk} = \beta_0 + \beta_l^3 + \cdots$$

– Si deux variables sont explicatives, par exemple X^2 et X^3, le modèle doit conserver les termes :

$$\eta_{jk} = \beta_0 + \beta_k^2 + \beta_l^3 + \beta_{kl}^{23} + \cdots$$

La généralisation à plus de trois variables ne pose pas de problème théorique. Les difficultés viennent de l'explosion combinatoire du nombre de termes d'interaction et de la complexité des structures d'indépendance. D'autre part, si le nombre de variables est grand, on est souvent confronté à des tables de contingence creuses (beaucoup de cellules vides) qui rendent défaillant le modèle log-linéaire. Une étude exploratoire (correspondances multiples par exemple) préalable est nécessaire afin de réduire le nombre des variables considérées et celui de leurs modalités.

3 Choix de modèle

3.1 Recherche pas à pas

Principalement deux critères (test du rapport de vraisemblance et test de Wald), décrits en annexe pour un cadre plus général, sont considérés. Ces critères sont utilisés comme le test de Fisher du modèle linéaire gaussien. Ils permettent de comparer un modèle avec un sous-modèle et d'évaluer l'intérêt de la présence des termes complémentaires. On suit ainsi une stratégie descendante à partir du modèle complet ou saturé dans le cas du modèle log-linéaire. L'idée est de supprimer, un terme à la fois, la composante d'interaction ou l'effet principal qui apparaît comme le moins significatif au sens du rapport de vraisemblance ou du test de Wald. Les tests présentent une structure hiérarchisée. SAS facilite cette recherche en produisant une décomposition (Type III) de ces indices permettant de comparer chacun des sous-modèles excluant un des termes avec le modèle les incluant tous.

Attention, du fait de l'utilisation d'une transformation non linéaire (log), même si des facteurs sont orthogonaux, aucune propriété d'orthogonalité ne peut être prise en compte pour l'étude des hypothèses. Ceci impose l'élimination des termes un par un et la ré-estimation du modèle. D'autre part, un terme principal ne peut être supprimé que s'il n'intervient plus dans des termes d'interaction. Enfin, selon les conditions expérimentales qui peuvent fixer les marges d'une table de contingence, la présence de certains paramètres est imposée dans un modèle log-linéaire.

4 Exemples

4.1 Modèle poissonien

On étudie les résultats d'une étude préalable à la législation sur le port de la ceinture de sécurité dans la province de l'Alberta à Edmonton au Canada (Jobson, 1991). Un échantillon de 86 769 rapports d'accidents de voitures ont été compulsés afin d'extraire une table croisant :

1. Etat du conducteur : Normal ou Alcoolisé
2. Port de la ceinture : Oui Non
3. Gravité des blessures : 0 : rien à 3 : fatales

La procédure `genmod` est utilisée :

```
proc genmod data=sasuser.ceinture;
class co ce b ;
model effectif=co|ce|b @2  /type3 obstats dist=poisson;
run;
```

Une extraction des résultats donnent :

```
            Criteria For Assessing Goodness Of Fit
          Criterion             DF         Value      Value/DF
          Deviance               3        5.0136        1.6712

                 LR Statistics For Type 3 Analysis
                 Source       DF   ChiSquare  Pr>Chi

                 CO            1   3431.0877  0.0001
                 CE            1   3041.5499  0.0001
                 CO*CE         1    377.0042  0.0001
                 B             3  28282.8778  0.0001
                 CO*B          3    474.7162  0.0001
                 CE*B          3     42.3170  0.0001

                   Analysis Of Parameter Estimates
    Parameter           DF    Estimate     Std Err   ChiSquare  Pr>Chi
    INTERCEPT            1      3.6341      0.1550    550.0570  0.0001
    CO         A         1     -2.2152      0.1438    237.3628  0.0001
    CE         N         1      1.8345      0.1655    122.8289  0.0001
    CO*CE      A  N      1      0.9343      0.0545    293.9236  0.0001
    B          0         1      5.7991      0.1552   1396.7752  0.0001
    B          1         1      2.7848      0.1598    303.6298  0.0001
    B          2         1      2.1884      0.1637    178.7983  0.0001
    CO*B       A  0      1     -1.4622      0.1354    116.5900  0.0001
    CO*B       A  1      1     -0.6872      0.1423     23.3154  0.0001
    CO*B       A  2      1     -0.5535      0.1452     14.5293  0.0001
    CE*B       N  0      1     -0.2333      0.1658      1.9807  0.1593
    CE*B       N  1      1     -0.0902      0.1708      0.2786  0.5976
    CE*B       N  2      1      0.0741      0.1748      0.1799  0.6715

                       Observation Statistics
EFFECTIF       Pred      Xbeta        Std     HessWgt      Lower      Upper
   12500      12497     9.4332   0.008930       12497      12280      12718
     604   613.3370     6.4189     0.0395   613.3370   567.6707   662.6770
     344   337.8089     5.8225     0.0530   337.8089   304.5010   374.7601
      38    37.8677     3.6341     0.1550    37.8677    27.9495    51.3053
   61971      61974    11.0345   0.004016      61974      61488      62464
   ...
```

Les résultats montrent que le modèle de toute interaction d'ordre 2 est acceptable (déviance) et il semble que tous les termes soient nécessaires, toutes les interactions doivent être présentes au sens du test de Wald.