Le modèle log-linéaire

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)

HAX912X

1. Introduction

2. Le modèle log-linéaire
   - Modèle log-linéaire pour des tableaux $\mathrm{I}_1 \times \mathrm{I}_2$
   - Définition des modèles log-linéaires
   - Estimation et inférence
   - Plan de Poisson

3. Modèles graphiques d'association
   - Notations et définitions
   - Graphe d'association
   - Les modèles graphiques d'association
   - Choix de modèles graphiques
   - Mise en oeuvre sous R

Introduction

Objectif : décrire les relations qui existent entre un certain nombre de variables qualitatives

Ce modèle a la particularité de ne pas nécessiter a priori de distinction entre la variable à expliquer et les variables explicatives

On ne parle plus d'un modèle de régression mais d'un modèle d'association

Introduction

On considère K variables qualitatives, facteurs, la k-ième variable x_k a I_k modalités notées $\{1, \ldots, I_k\}$

On peut construire une table de contingence ayant pour effectif

$$n_{i_1, i_2, \ldots, i_K}$$

où $i_k \in \{1, \ldots, I_k\}$ et $k \in \{1, \ldots, K\}$

Introduction

Objectif : expliquer les logarithmes des valeurs attendues des effectifs à l'aide des niveaux des facteurs et des interactions entre ces niveaux

Le modèle log-linéaire

K facteurs avec modalités $I_1, \ldots, I_K$
Nombre total de cellules : $J = I_1 \times \ldots \times I_K$
j correspond à une cellule $i_1, \ldots, i_k$

Soit π_j la probabilité que l'individu tombe dans la cellule j de la table de contingence

On considère n individus indépendants, la variable n_j correspondant à l'effectif de la cellule j est telle que :

$$n_j \sim \mathscr{B}(n, \pi_j)$$

Le modèle log-linéaire

Nous avons $\mathbb{E}(n_j) = n\pi_j$ et $\sum_j \mathbb{E}(n_j) = n$

Le vecteur aléatoire $(n_1, n_2, \ldots, n_J)$ suit une distribution multinomiale de paramètres $(n, (\pi_1, \ldots, \pi_J))$

Les probabilités $\pi_1, \ldots, \pi_J$ sont les paramètres inconnus du modèle qui caractérisent complètement la loi multinomiale

Si on impose aucune contrainte, on parle du modèle saturé, autant de paramètres que d'observations

Le modèle log-linéaire

On note $\mu_j = n\pi_j$

Pour le modèle saturé, on estime μ_j par $\hat{\mu}_j = n_j$ ce qui implique que $\hat{\pi}_j = \frac{n_j}{n}$

On remarque que $\sum_j \hat{\pi}_j = 1$

Il ne reste plus aucun degré de liberté ; afin de retrouver une description plus parcimonieuse de nos données, il est classique d'imposer une structure sur les paramètres inconnus

La tâche d'une analyse log-linéaire est de chercher une telle structure et de l'interpréter

Le modèle log-linéaire
Modèle log-linéaire pour des tableaux $I_1 \times I_2$

On considère ici seulement deux variables qualitatives (facteurs)

Le modèle d'indépendance s'écrit

$$\log(\mu_j) = \log(\mu_{i_1,i_2}) = \mu + \alpha_{1,i_1} + \alpha_{2,i_2}$$

avec $\sum_{i_1} \alpha_{1,i_1} = \sum_{i_2} \alpha_{2,i_2} = 0$

Le modèle log-linéaire
Modèle log-linéaire pour des tableaux $I_1 \times I_2$

Nous avons

$$\mathbb{E}(n_j) = \mathbb{E}(n_{i_1,i_2}) = \exp(\mu + \alpha_{1,i_1} + \alpha_{2,i_2})$$

Sous cette hypothèse, les facteurs x_1 et x_2 sous-jacents à la table de contingence sont indépendants

Le modèle log-linéaire
Modèle log-linéaire pour des tableaux $I_1 \times I_2$

Le modèle nul s'écrit

$$\log(\mu_j) = \log(\mu_{i_1,i_2}) = \mu$$

Nous avons :

$$\mathbb{E}(n_j) = \mathbb{E}(n_{i_1,i_2}) = \exp(\mu)$$

Le modèle log-linéaire
Modèle log-linéaire pour des tableaux $I_1 \times I_2$

Un modèle possible s'écrit

$$\log(\mu_j) = \log(\mu_{i_1,i_2}) = \mu + \alpha_{1,i_1}$$

avec $\sum_{i_1} \alpha_{1,i_1} = 0$

Nous avons :

$$\mathbb{E}(n_j) = \mathbb{E}(n_{i_1,i_2}) = \exp(\mu + \alpha_{1,i_1})$$

Le modèle log-linéaire
Modèle log-linéaire pour des tableaux $I_1 \times I_2$

Le modèle saturé s'écrit

$$\log(\mu_j) = \log(\mu_{i_1,i_2}) = \mu + \alpha_{1,i_1} + \alpha_{2,i_2} + \alpha_{12,i_1,i_2}$$

avec $\sum_{i_1} \alpha_{1,i_1} = \sum_{i_2} \alpha_{2,i_2} = \sum_{i_1} \alpha_{12,i_1,i_2} = \sum_{i_2} \alpha_{12,i_1,i_2} = 0$

Nombre de paramètres : $I_1 \times I_2 - 1$

Le modèle log-linéaire
Définition des modèles log-linéaires

Pour simplifier les notations, nous avons aligné tous les éléments du tableau dans un long vecteur $(n_1, \ldots, n_J)$ où J est le nombre de cellule du tableau et les n_i sont les effectifs de chaque cellule

On note $\mu_j = \mathbb{E}(n_j) = n\pi_j$ ($\sum_{j=1}^{J} \mu_j = n$) et $\mu = (\mu_1, \ldots, \mu_J)$

Le modèle log-linéaire

Définition des modèles log-linéaires

Le modèle log-linéaire impose que

$$\log(\mu) = X\theta$$

où θ est un paramètre inconnu de dimension p

La matrice X de dimension $J \times p$ contient des variables indicatrices spécifiant les niveaux et les interactions

$(n_1, \ldots, n_J)$ suit une loi-multinomiale

$$f(n_1, \ldots, n_J) = \frac{J!}{\prod_{j=1}^{J} n_j!} \prod_{j=1}^{J} \pi_j^{n_j}$$

Le modèle log-linéaire
Estimation et inférence

On veut estimer θ sous la contrainte que $\sum_{j=1}^{J} \mathbb{E}(n_j) = n$

On utilise la méthode du maximum de vraisemblance

$$LV(\theta; (n_1, \ldots, n_J)) = \sum_{j=1}^{J} n_j \log(\mu_j) + \text{cst}$$

$$LV(\theta; (n_1, \ldots, n_J)) = \sum_{j=1}^{J} n_j \sum_{k=1}^{p} x_{jk}\theta_k + \text{cst}$$

Maximiser LV sous contrainte $\sum_{j=1}^{J} \exp\left(\sum_{k=1}^{p} x_{jk}\theta_j\right) = n$ exige l'utilisation d'un algorithme numérique

Le modèle log-linéaire
Plan de Poisson

Une hypothèse de départ était que le vecteur des effectifs suit une distribution multinomiale

Si la taille de l'échantillon n'est pas décidée à l'avance, cela ne convient pas

Si n est aléatoire, le vecteur des effectifs ne suit plus une loi multinomiale

C'est par exemple le cas lorsque l'on obtient les valeurs des facteurs pendant une durée pré-déterminée

Le modèle log-linéaire
Plan de Poisson

Dans ce cas, on peut utiliser une loi de Poisson

$$f(n_1, \ldots, n_J) = \prod_{j=1}^{J} \frac{\mu_j^{n_j} \exp(-\mu_j)}{n_j!}$$

Avec $\log(\mu_j) = \sum_{k=1}^{p} X_{jk}\theta_k$, on obtient

$$LV(\theta; (n_1, \ldots, n_J)) = \sum_{j=1}^{J} n_j \sum_{k=1}^{p} X_{jk}\theta_k - \sum_{j=1}^{J} \exp\left(\sum_{k=1}^{p} X_{jk}\theta_k\right) + \text{cst}$$

On maximise LV sans contrainte, pas de solution explicite

Modèles graphiques d'association

Un modèle graphique d'association est un modèle log-linéaire qui spécifie de façon unique les relations de dépendance par un graphe non orienté

Modèles graphiques d'association
Notations et définitions

Notons $A, B, C, \ldots$ les facteurs qui nous concernent, dans l'écriture d'un modèle log-linéaire :

- A représente les effets principaux associés à A (interaction d'ordre 0)
- $A : B$ représente tous les termes d'interactions entre les niveaux de A et de B (interaction d'ordre 1)
- $A * B = A + B + A : B$
- $A : B : C$ représente tous les termes d'interactions entre A, B et C (interaction d'ordre 2)
- $A * B * C = A + B + C + A : B + A : C + B : C + A : B : C$

Modèles graphiques d'association
Notations et définitions

Nous nous limitons aux modèles hiérarchiques : si un terme d'interaction d'ordre k est présent, toutes les interactions d'ordre inférieurs y sont aussi

Dans ce cas, il est toujours possible d'écrire l'équation du modèle en utilisant les opérateurs $+$ ou $*$

L'équation du modèle détermine alors les relations d'indépendance conditionnelles entre les variables

Modèles graphiques d'association
Graphe d'association

Deux règles de base :

- chaque sommet représente un facteur présent dans le modèle ;
- l'arrête entre les sommets $\mathcal{A}$ et $\mathcal{B}$ existe ssi il y a une interaction d'ordre 1 entre $\mathcal{A}$ et $\mathcal{B}$ dans le modèle

Modèles graphiques d'association
Graphe d'association

- deux sommets A et B sont dits connectés s'il existe un chemin de A vers B
- si un sous-ensemble de sommets S est tel que chaque chemin de A vers B passe par S, on dit que S sépare A et B

Théorème de séparation

1) A est indépendant de B ssi A et B ne sont pas connectés
2) A est indépendant de B sachant S ssi A et B sont séparés par S

Modèles graphiques d'association
Les modèles graphiques d'association

Soit $\mathcal{G}(\mathcal{M})$ le graphe d'association du modèle hiérarchique $\mathcal{M}$

Un modèle graphique d'association $\mathcal{M}_{\mathcal{G}}$ est défini comme étant un modèle hiérarchique $\mathcal{M}$ tel que $\mathcal{G}(\mathcal{M}) = \mathcal{G}(\mathcal{M}_{\mathcal{G}})$

Chaque modèle hiérarchique $\mathcal{M}$ avec $\mathcal{G}(\mathcal{M}) = \mathcal{G}(\mathcal{M}_{\mathcal{G}})$ doit être un sous-modèle de $\mathcal{M}_{\mathcal{G}}$

Modèles graphiques d'association
Les modèles graphiques d'association

À partir d'un graphe $\mathscr{G}_0$ donné, il n'est pas difficile de trouver le modèle graphique d'association

On cherche les cliques de taille maximale du graphe ; une clique est un ensemble de sommets tous connectés deux à deux

Les cliques constituent les composantes de $\mathscr{M}_{\mathscr{G}}$

Modèles graphiques d'association
Les modèles graphiques d'association

L'ensemble des modèles log-linéaires graphiques est une sous-ensemble des modèles log-linéaires hiérarchiques

Seules les cliques du graphe doivent apparaitre dans l'équation hérarchique définissant le modèle à l'aide des opérateurs $+$ et $*$

Modèles graphiques d'association
Choix de modèles graphiques

Pour d facteurs, il y a $\frac{d(d-1)}{2}$ arrêtes possibles et $2^{\frac{d(d-1)}{2}}$ graphes possibles

On ne peut pas analyser tous les modèles, on va utiliser une procédure de type backward

On part du graphe complet et de nouveaux graphes sont crées en éliminant certaines arrêtes : on supprime successivement la connexion pour laquelle le fait de l'enlever fait baisser le plus le critère considéré (AIC ou BIC)

On utilisera plutôt le critère BIC

Modèles graphiques d'association
Mise en oeuvre sous R

La fonction `dmod` de la bibliothèque `gRim` permet d'estimer θ et de choisir le meilleur modèle