Chapitre 3

LE MODÈLE LOG-LINÉAIRE

C. Croux

3.1 Introduction

Le modèle linéaire usuel essaie de prédire ou d'expliquer une variable Y, mesurée sur une échelle continue, à partir d'un ensemble de K variables explicatives $X_1, \ldots, X_K$. Ces variables explicatives peuvent être continues ou catégorielles. Dans ce dernier cas, une variable à J modalités nécessitera l'insertion de $J-1$ variables indicatrices (ou variables « dummies ») dans le membre de droite de l'équation de régression. Si toutes les variables explicatives sont qualitatives, on sera plutôt dans le cas de ce que l'on appelle un modèle d'analyse de la variance (ANOVA). Ce type de modèle est, dans le sens strict, un cas particulier du modèle de régression classique. Les modèles ANOVA sont traités séparément dans la littérature statistique à cause des problèmes spécifiques qu'ils engendrent. L'apparition de termes d'interaction, qui mesurent le degré avec lequel les variables explicatives jouent ensemble, est typique pour ces derniers.

Si la variable à expliquer n'est plus continue, mais qualitative ou discrète, le modèle linéaire n'est plus approprié. D'autres modèles, comme le modèle de régression logistique ou le modèle de Poisson, sont plus adaptés. La théorie des modèles linéaires généralisés donne un cadre général pour traiter de tels cas et sera étudiée dans un chapitre ultérieur. Le modèle log-linéaire qui sera introduit ici est un exemple de modèle linéaire généralisé.

L'usage du modèle log-linéaire est approprié pour rechercher des relations entre un certain nombre de variables qualitatives. Celui-ci a la particularité de ne pas nécessiter, a priori, de distinction entre la variable à expliquer et les variables explicatives. Pour cela, on ne parlera plus d'un modèle de régression, mais d'un modèle d'association.

Par analogie avec l'analyse de la variance, les K variables qualitatives en question seront appelées les *facteurs*. Le k-ième facteur peut prendre un total de I_k modalités ou niveaux. A partir de ces K facteurs, le tableau de contingence (à K entrées) est construit, ayant comme effectifs les nombres :

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

où chaque i_k varie entre 1 et I_k, pour $k = 1, \ldots, K$. Un tableau de contingence $I_1 \times I_2 \times \ldots \times I_K$ est ainsi obtenu. L'idée est maintenant d'expliquer les logarithmes des valeurs attendues des effectifs à l'aide des niveaux correspondants des facteurs et des interactions entre ces niveaux.

La formulation du modèle log-linéaire sera donnée dans la section suivante, ainsi que certains aspects d'inférence statistique et le critère de déviance. La section 3 donne un exemple et la section 4 discute différentes options possibles pour la sélection du modèle final. Dans cette section, la manière d'affronter une analyse résiduelle sera également décrite. Le lien entre le modèle log-linéaire et le modèle logit est établi en section 5. La section 6 va servir à montrer comment des graphes d'association peuvent aider à interpréter le résultat d'une analyse log-linéaire. Finalement, la dernière section fera le lien avec l'analyse de correspondance.

3.2 Le modèle log-linéaire

Supposons qu'ont été collectées, pour n individus, les valeurs de K facteurs. Nous avons par exemple demandé, à $n = 1000$ personnes, sélectionnées indépendamment et aléatoirement dans la population des automobilistes ; (A) s'ils utilisent leur voiture pour la plupart des trajets pour des raisons professionnelles (oui/non) (B) s'ils respectent toujours la vitesse maximale (oui/non), (C) leur sexe (hommes/femmes), et (D) l'âge de leur voiture (< 2 ans, entre 2 et 5 ans, > 5 ans). Les réponses sont insérées dans un tableau de contingence $2 \times 2 \times 2 \times 3$, qui est donc de dimension 4.

Soit $\pi_{i_1,\ldots,i_K}$ la probabilité qu'un individu tombe dans la cellule $(i_1, \ldots, i_K)$ de la table de contingence. Définissons les valeurs attendues des variables aléatoires associées aux effectifs $N_{i_1,\ldots,i_K}$ par $\mu_{i_1,\ \ ,i_K}$:

$$E(N_{i_1,\ldots,i_K}) := \mu_{i_1,\ldots,i_K}. \tag{3.1}$$

La façon de recueillir les données implique que le vecteur des effectifs $n_{i_1,\ldots,i_K}$ suit une distribution multinomiale de paramètres n et $\{\pi_{i_1,i_2,\ldots,i_K}; 1 \leq i_k \leq I_k, 1 \leq k \leq K\}$. Il s'en suit que :

$$\mu_{i_1,\ldots,i_K} = n\pi_{i_1,i_2,\ldots,i_K}, \tag{3.2}$$

et

$$\sum_{i_1,i_2,\ldots,i_K} \mu_{i_1,\ldots,i_K} = n. \tag{3.3}$$

Les $\mu_{i_1,\ldots,i_K}$ sont les paramètres inconnus du modèle, qui caractérisent complètement la distribution multinomiale. Leurs valeurs estimées $\hat{\mu}_{i_1,\ .,i_K}$ donnent les *valeurs ajustées* du tableau de contingence. Ces paramètres sont estimés ici d'une façon naturelle (et d'ailleurs aussi par le principe de maximum de vraisemblance) comme :

$$\hat{\mu}_{i_1,\ .,i_K} = n_{i_1,\ldots,i_K}. \tag{3.4}$$

Si le modèle n'impose aucune contrainte sur les valeurs des paramètres, on parlera d'un *modèle saturé*. Dans un modèle de ce type, on a un ajustement parfait des effectifs du tableau, comme le montre l'équation (3.4). Par contre, le nombre total de paramètres $\mu_{i_1,\ldots,i_K}$ étant le même que le nombre d'effectifs dans le tableau de contingence, il ne reste plus aucun degré de liberté.

Afin de retrouver une description plus parcimonieuse de nos données et d'augmenter le nombre de degrés de liberté (et par conséquent la précision statistique des estimateurs), il est indispensable d'imposer une structure sur les paramètres inconnus. La tâche d'une analyse log-linéaire est de chercher une telle structure et d'essayer de l'interpréter.

3.2.1 Modèles log-linéaires pour des tableaux $I \times J$

Le modèle d'indépendance pour une table $I \times J$

Ce modèle s'écrit :

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j, \tag{3.5}$$

pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$. Le paramètre μ donne l'effet général, le paramètre α_i l'effet ligne, et le paramètre β_j l'effet colonne. Afin d'identifier tous les paramètres, on ajoute les contraintes :

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0.$$

L'analogie avec un modèle ANOVA à 2 facteurs (et sans interaction) est évidente.

Il est intéressant d'illustrer la notation matricielle de l'équation (3.5) dans le cas, par exemple, où $I = 2, J = 3$. Le vecteur des paramètres inconnus est donné par $\boldsymbol{\theta} = (\mu, \alpha_1, \beta_1, \beta_2)'$, étant donné que $\alpha_2 = -\alpha_1$ et $\beta_3 = -\beta_1 - \beta_2$.

Le modèle se réécrit comme :

$$\begin{pmatrix} \log(\mu_{11}) \\ \log(\mu_{12}) \\ \log(\mu_{13}) \\ \log(\mu_{21}) \\ \log(\mu_{22}) \\ \log(\mu_{23}) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Il existe donc une matrice $\mathbf{X}$ telle que :

$$\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\theta},$$

où $\boldsymbol{\mu}$ est le vecteur des espérances des effectifs du tableau de contingence.

On peut facilement montrer (cf. Christensen [1990, page 49]) que les distributions marginales $\pi_{i.}$ et $\pi_{.j}$ sont indépendantes, si et seulement si l'équation (3.5) est vérifiée. Autrement dit :

$$\pi_{ij} = \pi_{i.}\pi_{.j} \Leftrightarrow \exists\, \alpha_i, \beta_j \text{ et } \mu \text{ tels que } \log(\mu_{ij}) = \mu + \alpha_i + \beta_j.$$

L'équation du modèle impose, par conséquence, l'indépendance entre les lignes et colonnes du tableau de contingence.

Le modèle saturé pour une table $I \times J$

Ce modèle s'écrit

$$\log(\mu_{ij}) = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \tag{3.6}$$

pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$. Le paramètre $(\alpha\beta)_{ij}$ est appelé une interaction. Il est indispensable, à présent, d'ajouter les contraintes :

$$\sum_{i=1}^{I} (\alpha\beta)_{ij} = \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0.$$

L'analogie avec un modèle ANOVA à 2 facteurs (et muni d'interactions) est claire. Lorsque $I = 2$ et $J = 3$ le vecteur des paramètres inconnus est $\boldsymbol{\theta} = (\mu, \alpha_1, \beta_1, \beta_2, (\alpha\beta)_{11}, (\alpha\beta)_{12})'$ qui est de la même dimension que $\boldsymbol{\mu}$. Vu que le nombre de paramètres indépendants est équivalent au nombre de cellules, on aura un ajustement parfait du tableau de contingence et donc un modèle saturé. La notation matricielle de (3.6) est donnée par :

$$\begin{pmatrix} \log(\mu_{11}) \\ \log(\mu_{12}) \\ \log(\mu_{13}) \\ \log(\mu_{21}) \\ \log(\mu_{22}) \\ \log(\mu_{23}) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \end{pmatrix},$$

ce qui s'écrit en abrégé comme $\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\theta}$. D'autres paramétrisations linéaires du modèle sont possibles. Autrement dit, on peut encore trouver une autre matrice $\mathbf{X}^*$ et un autre vecteur $\boldsymbol{\theta}^*$ tels que $\log(\boldsymbol{\mu}) = \mathbf{X}^*\boldsymbol{\theta}^*$

D'autres modèles pour une table $I \times J$

Le modèle d'indépendance et le modèle saturé sont sans aucun doute les plus utilisés pour l'étude d'association dans des tableaux à 2 entrées. Ce ne sont cependant pas les seuls modèles log-linéaires. Si l'hypothèse $H_0 : \alpha_1 = \ldots = \alpha_I (= 0)$ dans un modèle d'indépendance n'est pas rejetée, on peut spécifier le modèle comme étant :

$$\log(\mu_{ij}) = \mu + \beta_j.$$

On en déduit facilement que π_{ij} ne dépend plus de l'indice i, ce qui reste vrai pour pour $\pi_{i.}$. Il vient $\pi_{i.} = 1/I$, et donc la distribution marginale des lignes est équiprobable ou uniforme.

Après ces exemples introductifs, nous pouvons à présent présenter une formulation plus générale du modèle log-linéaire.

3.2.2 Définition du modèle log-linéaire

Pour simplifier la notation, alignons tous les éléments du tableau de contingence dans un long vecteur $Y = (Y_1, \ldots, Y_N)$, où N est le nombre de cellules du tableau et où les Y_i représentent les effectifs. On impose que les logarithmes des valeurs attendues des effectifs dépendent d'une façon linéaire de certaines variables explicatives. Soit $\mathbf{X}$ la matrice des observations des variables explicatives et $\boldsymbol{\mu} = E[Y]$, on suppose alors que :

$$\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\theta}, \tag{3.7}$$

où $\boldsymbol{\theta}$ est un paramètre inconnu de dimension p inférieure ou égale à N. Dans le cadre d'une modélisation log-linéaire, les colonnes de X contiennent des variables indicatrices spécifiant les niveaux et les interactions[1].

Le problème d'estimation se réduit à estimer le paramètre $\boldsymbol{\theta}$, ce qui donnera, via la relation (3.7), une estimation pour $\boldsymbol{\mu}$. Le paramètre $\boldsymbol{\theta}$ sera estimé par maximum de vraisemblance, sous la contrainte que :

$$\sum_{i=1}^{N} \mu_i = n, \tag{3.8}$$

où n est la taille de l'échantillon et N le nombre de cellules.

[1] Si un facteur est ordinal, on peut attacher aux différentes modalités un score. Dans ce cas, la colonne correspondante en X contiendra les scores. Pour la modélisation des variables qualitatives ordinales, cf. Agresti [1984].

3.2.3 Estimation et inférence

Comme dit précédemment, le vecteur $(Y_1, \ldots, Y_N)$ suit une distribution multinomiale. Sa densité est donc égale à :

$$f(y_1, \ldots, y_N) = \frac{N!}{\prod_{i=1}^{N} y_i!} \prod_{i=1}^{N} \pi_i^{y_i},$$

où π_i est la probabilité d'appartenir à la cellule i. Vu que $n\pi_i = E[Y_i] = \mu_i$, la fonction de log-vraisemblance est donnée par :

$$l(\mu_1, \ldots, \mu_N) = \sum_{i=1}^{N} y_i \log \mu_i + c, \tag{3.9}$$

où c ne dépend pas des paramètres inconnus. En exploitant (3.7), l'équation devient

$$l(\mu_1, \ldots, \mu_N) = \sum_{j=1}^{p} \left(\sum_{i=1}^{N} y_i \mathbf{X}_{ij} \right) \theta_j + c, \tag{3.10}$$

où $\dim(\boldsymbol{\theta}) = p$. L'équation (3.10) permet de voir que la distribution de $(Y_1, \ldots, Y_N)$ est un membre de la famille des distributions exponentielles. La théorie de l'inférence statistique pour les familles exponentielles peut donc être utilisée (cf. Andersen [1997, page 29]). Remarquons cependant que la dimension de la famille exponentielle (3.10) n'est pas p, mais $p - 1$, car il est encore nécessaire d'incorporer la condition (3.8) dans la vraisemblance. La maximisation de (3.10) exige l'emploi d'un algorithme numérique (sauf pour quelques cas simples, comme les exemples vus pour les tables de dimension 2). Un algorithme qui mérite d'être cité est le *Iterative proportional fitting algorithm* de Deming et Stephan [1940].

Une hypothèse de départ était que les effectifs suivent une distribution multinomiale simple. Cette hypothèse convient pour la plupart des exemples pratiques, mais il s'avère important de considérer encore deux autres plans d'échantillonnage.

3.2.4 Autres plans d'échantillonnage

Produits des distributions multinomiales

Ici, la population est stratifiée suivant les modalités d'un facteur du modèle, ce qui donne les strates $S_1, S_2, \ldots, S_m$. Chaque strate S_j a n_j éléments et les strates forment une partition de $\{1, \ldots, n\}$, donc $\sum_j n_j = n$. La taille totale de l'échantillon, n, est fixée a priori, ainsi que la taille des différentes strates.

Dans notre exemple avec les automobilistes, la stratification de la population pourrait être suivant le sexe, et 500 personnes tirées de la population des

automobilistes seraient féminines et 500 masculines. Les questions (A), (B) et (D) auraient pu être posées à ces deux groupes de 500 personnes. On serait ainsi dans le cas d'une stratification suivant les modalités d'un facteur du modèle.

La fonction de vraisemblance est donnée, dans ce cas, par un produit de lois multinomiales :

$$f(y_1, \ldots, y_N) = \prod_{j=1}^{m} \frac{n_j!}{\prod_{i \in S_j} y_i!} \prod_{i \in S_j} \pi_i^{y_i},$$

où :

$$\mu_i := E[y_i] = \pi_i n_j \quad \text{pour } i \in S_j.$$

Il est aisé de vérifier que la log-vraisemblance est la suivante :

$$l(\mu_1, \ldots, \mu_N) = \sum_{i=1}^{N} y_i \log \mu_i + \tilde{c}, \tag{3.11}$$

où $\tilde{c}$ ne dépend pas des inconnues $\mu_1, \ldots, \mu_N$. On obtient, à une constante près, la même fonction de log-vraisemblance que pour le plan d'échantillonnage multinomial. La maximisation de (3.11) est maintenant effectuée sous les contraintes :

$$\sum_{i \in S_j} \mu_i = n_j \quad j = 1, \ldots m. \tag{3.12}$$

Pour presque tous les modèles rencontrés, l'estimateur pour un plan d'échantillonnage multinomial simple vérifiera aussi la condition (3.12), ce qui implique que cet estimateur sera identique à l'estimateur du maximum de vraisemblance pour un produit des distributions multinomiales. En pratique, la distinction entre ces deux plans d'échantillonnage n'est pas faite (Everitt and Dunn, [1991, page 171]).

Plan de Poisson

Dans ce cas-ci, la taille de l'échantillon n'est pas décidée à l'avance. Pendant une certaine période de durée prédéterminée on observe, pour une série d'observations, les valeurs des facteurs. Par exemple, on interroge des automobilistes pendant une période de 6 heures. L'échantillon sera constitué par les automobilistes interviewés pendant cette période et sa taille sera aléatoire. Supposons maintenant que le nombre total des gens qui ont répondu une combinaison (i_1, i_2, i_3, i_4) à l'ensemble des 4 questions suive une distribution de Poisson de paramètre (et moyenne) μ_i. Alors :

$$f(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

et la log-vraisemblance est donnée par :

$$l(\mu_1, \ldots, \mu_N) = \sum_{i=1}^{N} y_i \log \mu_i - \sum_{i=1}^{N} \mu_i + c. \tag{3.13}$$

Soient x_i les lignes de la matrice $\mathbf{X}$, la condition du premier ordre pour l'estimateur du maximum de vraisemblance est donnée par (cf. Gourieroux, [1989, page 299], ou Chapitre 9 de cet ouvrage) :

$$\sum_{i=1}^{N} x_i(y_i - \mu_i) = 0.$$

Il s'en suit que l'estimateur de maximum de vraisemblance doit satisfaire :

$$\sum_i y_i = \sum_i \mu_i,$$

dès qu'une certaine combinaison linéaire des colonnes de $\mathbf{X}$ est constante (ce qui sera toujours le cas chez nous). On en conclut que l'estimateur du maximum de vraisemblance doit maximiser :

$$l(\mu_1, \ldots, \mu_N) = \sum_{i=1}^{N} y_i \log \mu_i + \bar{c}$$

ce qui est de la même forme que (3.9).

Malgré le fait qu'un plan de Poisson semble peu applicable en pratique, il sera tout de même utilisé comme une sorte d'approximation de la vraie distribution. On parle d'un « surrogate Poisson model ». Cette approche est correcte dans le sens que les estimateurs et les erreurs-types seront identiques dans un modèle de Poisson et dans les modèles multinomiaux (sous de légères conditions sur la matrice $\mathbf{X}$). En plus, il est connu que la distribution de Poisson, conditionnellement à la taille de l'échantillon, est identique à une distribution multinomiale. L'avantage d'utiliser le modèle de Poisson est qu'il s'agit d'un modèle linéaire généralisé bien connu et implémenté dans presque tous les logiciels statistiques.

3.2.5 Critère de déviance

La qualité de l'ajustement des cellules du tableau de contingence par un modèle log-linéaire M_A peut être mesurée par la statistique G^2, définie par

$$G^2 = 2 \sum_{i=1}^{N} y_i \log \frac{y_i}{\hat{\mu}_i}. \tag{3.14}$$

Soit $\log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\theta}$, l'équation du modèle M_A. L'équation (3.14) est deux fois la différence entre les maxima de la fonction de log-vraisemblance du modèle saturé et du modèle M_A (ceci est vrai pour les 3 plans d'échantillonnage considérés). Le critère de déviance n'est donc rien d'autre que la statistique du test du rapport de maximum de vraisemblance, qui teste l'hypothèse nulle $H_0 : \log(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\theta}$. Le critère G^2 est appelé la déviance du modèle, et son rôle crucial dans les procédures de sélection d'un modèle sera montré dans la suite. On sait, par la théorie des modèles linéaires généralisés, que G^2 suit asymptotiquement, et sous H_0, une distribution du khi-carré avec un nombre de degrés de liberté donné par :

df=(le nombre de cellules)-(le nombre de paramètres indépendamment estimés).

Si la statistique G^2 reste inférieure à la valeur critique d'un khi-carré à df degrés de liberté, le modèle ne sera pas rejeté.

Ce critère de déviance est également utile pour comparer différents modèles. Supposons que M_A est un sous-modèle de M_B, et nous désirons tester si M_B donne lieu à un ajustement de la table de contingence significativement meilleur. Notons $G^2(A)$ et $G^2(B)$ les deux critères de déviance. La statistique du rapport de maximum de vraisemblance utilisée sera donnée par :

$$G^2(B|A) = 2\sum_{i=1}^{n} y_i \log(\frac{\hat{\mu}_i^B}{\hat{\mu}_i^A}),$$

où $\hat{\mu}_i^A$ et $\hat{\mu}_i^B$ seront respectivement les valeurs ajustées par M_A et M_B. Il en résulte que :

$$G^2(B|A) = G^2(A) - G^2(B).$$

La statistique de test $G^2(B|A)$ suit une distribution du khi-carré à df(M_A)-df(M_B) degrés de liberté. Si $G^2(B|A)$ est supérieur au quantile 0.95 de cette distribution, il sera préférable de continuer avec M_B. Dans le cas contraire, le plus petit modèle M_A sera plus approprié. Remarquons encore que le critère de déviance (3.14) compare un modèle M_A avec le modèle saturé M_S, car :

$$G^2(A) = G^2(S|A).$$

3.3 Exemple

Reprenons l'exemple des $n = 1000$ automobilistes et des 4 facteurs ; (A) usage professionnel de la voiture (oui/non) (B) respect de la vitesse maximale (oui/non), (C) sexe (hommes/femmes), et (D) âge de la voiture (< 2 ans, entre 2 et 5 ans, > 5 ans). Le tableau de contingence est de dimension 4 et contient $2 \times 2 \times 2 \times 3 = 24$ cellules. Comme premier essai, nous proposons un modèle

Tableau 3.1 – Fréquences observées pour toutes les combinaisons des facteurs A, B, C et D de l'exemple de 1000 automobilistes

A	B	C	D	Fréquence
oui	oui	M	< 2	51
non	oui	M	< 2	22
oui	non	M	< 2	70
non	non	M	< 2	12
oui	oui	F	< 2	52
non	oui	F	< 2	33
oui	non	F	< 2	20
non	non	F	< 2	11
oui	oui	M	$\geq 2, \leq 5$	63
non	oui	M	$\geq 2, \leq 5$	30
oui	non	M	$\geq 2, \leq 5$	115
non	non	M	$\geq 2, \leq 5$	43
oui	oui	F	$\geq 2, \leq 5$	19
non	oui	F	$\geq 2, \leq 5$	28
oui	non	F	$\geq 2, \leq 5$	47
non	non	F	$\geq 2, \leq 5$	32
oui	oui	M	> 5	70
non	oui	M	> 5	33
oui	non	M	> 5	47
non	non	M	> 5	25
oui	oui	F	> 5	46
non	oui	F	> 5	61
oui	non	F	> 5	32
non	non	F	> 5	38

Tableau 3.2 – Résultats de l'ajustement du modèle log-linéaire (3.15)

```
constante~: 3.606027

facteur A:       oui        non
                0.27      -0.27
facteur B:       oui        non
                0.06      -0.06
facteur C:         H          F
                0.11      -0.11
facteur D:       "<2"   ">=2,<=5"      ">5"
               -0.23       0.11        0.12

A X B:                 oui        non
           oui       -0.07       0.07
           non        0.07      -0.07
A X C:                   M          F
           oui         0.21      -0.21
           non        -0.21       0.21
B X C:                   M          F
           oui        -0.07       0.07
           non         0.07      -0.07
A  X D:                "<2"   ">=2,<=5"    ">5"
           oui         0.18      -0.04     -0.14
           non        -0.18       0.04      0.14
B X D:                 "<2"   ">=2,<=5"    ">5"
           oui         0.14      -0.28      0.14
           non        -0.14       0.28     -0.14
C X D:                 "<2"   ">=2,<=5"    ">5"
            M         -0.04       0.17     -0.13
            F          0.04      -0.17      0.13
```

log-linéaire avec des termes d'interaction du premier ordre. En utilisant des notations évidentes, notre modèle peut être écrit :

$$\log(\mu_{ijkl}) = \mu+\alpha_i+\beta_j+\gamma_k+\delta_l+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\alpha\delta)_{il}+(\beta\gamma)_{jk}+(\beta\delta)_{jl}+(\gamma\delta)_{jk}, \tag{3.15}$$

pour $i, j, k \in \{1, 2\}$ et $1 \leq l \leq 3$. Une notation encore plus courte pour cette équation est $A + B + C + D + A{:}B + A{:}C + A{:}D + B{:}C + B{:}D + C{:}D$.

Après avoir recueilli les données, la façon la plus simple pour représenter les effectifs est de les aligner dans un vecteur, comme pour le tableau 3.1 (données fictives). Nous obtenons ainsi une matrice des données qui est facile à manipuler. L'analyse statistique dans ce chapitre est faite avec le logiciel Splus et la liste des commandes que nous avons utilisées est reprise en Appendice. Nous avons créé le tableau de contingence de dimension 4, et ajusté le modèle (3.15). Ensuite, la déviance et la P-valeur associées ont été calculées.

Les estimations des effets principaux et des interactions sont données dans le tableau 3.2. Nous voyons qu'il y a plus de chauffeurs « professionnels », que la majorité ne respecte pas la vitesse maximale, et qu'une majorité des automobilistes est de sexe masculin. La catégorie des voitures neuves est minoritaire. L'étude des associations entre les facteurs est un sujet encore plus intéressant. Nous remarquons en effet une association (partielle) positive entre le fait d'être une femme et de respecter la vitesse maximale (cf. tableau B X C) et entre le fait d'avoir une nouvelle voiture et le fait de respecter la vitesse maximale. Les chiffres pris dans le tableau 3.2 mesurent des associations partielles, puisque toujours conditionnelles aux autres facteurs du modèle (cf. section 3.6). Ce ne sont cependant que des associations et nous ne pouvons encore rien conclure quant à leurs relations causales.

La déviance est de 19.87, ce qui donne pour le test du rapport de maximum de vraisemblance une P-valeur de 0.018. Nous en déduisons que le modèle est trop simple pour bien ajuster les effectifs de la table de contingence. Si l'on considère également toutes les interactions d' ordre 2, le modèle donne une déviance de 3.10 avec un $df = 2$ et une P-valeur associée de 0.211. (Cette P-valeur donne la probabilité qu'un khi-carré à 2 df soit plus grand que la déviance observée 3.10.) Le modèle n'est pas rejeté, mais est peut-être excessivement lourd. Lors de la section suivante, nous allons considérer le problème de sélection d'un modèle approprié. Nous reviendrons sur cet exemple à ce moment là.

3.4 Sélection du modèle

Le modèle saturé donne lieu à un ajustement parfait, mais il n'a plus aucun degré de liberté. En outre, sa complexité est excessive. Des modèles plus simples, avec seulement des effets principaux et pas de termes d'interaction, risquent, par contre, d'être incompatibles avec les données. Il faut donc une procédure de sélection.

3.4.1 Le critère AIC

Le choix entre deux modèles emboîtés (« nested ») peut se faire via un test de rapport de maximum de vraisemblance (comme expliqué dans la section 3.2.5). Un moyen alternatif, très populaire dans le contexte des modèles linéaires généralisés, est l'utilisation du critère d'information de Akaike (AIC). Pour un modèle M donné, celui-ci est défini comme étant :

$$\text{AIC} = G^2 - 2\text{df} + 2N, \tag{3.16}$$

où G^2 est la déviance du modèle et df le nombre de degrés de liberté. Le but est ici de rechercher le modèle qui minimise l'AIC, ou la différence $G^2 - 2\text{df}$ (car le nombre de cellules N est fixe). Bien que plusieurs variantes de la définition (3.16) de l'AIC existent, elles se basent toutes sur cette différence.

Pour arriver aux valeurs minimales de l'AIC, il faut tenir compte du fait qu'un modèle compliqué aura une valeur faible pour le premier terme (la déviance), mais sera pénalisé par un nombre de degrés de liberté moindre (le second terme). Le modèle « optimal » sera celui qui mènera à un bon compromis entre déviance et degrés de liberté.

Pour ce faire, il faut calculer ce critère pour tous les sous-modèles du modèle saturé. Si le nombre de facteurs et de modalités est élevé, ceci nous mènera à des calculs très lourds. D'autres procédures automatiques plus simples ont été développées. Parmi celles-ci, les plus connues sont la sélection rétrograde (« backward model selection ») et la sélection progressive (« forward model selection »).

3.4.2 Sélection pas-à-pas du modèle

En ce qui nous concerne nous allons utiliser une procédure pas-à-pas automatique qui combine à la fois des aspects d'une recherche progressive et rétrograde. La recherche couvrira l'ensemble des *modèles hiérarchiques,* respectant la propriété stipulant que : « *si un terme d'interaction d'ordre k entre k facteurs est présent dans le modèle, toutes les interactions d'ordre inférieur ou égal correspondant à ces facteurs doivent aussi y figurer* ».

Reprenons notre exemple avec les 1000 automobilistes, en commençant avec un modèle M_0, qui est un sous-modèle du modèle saturé. Un choix possible est de prendre M_0 : $A + B + C + D$, qui est une abréviation pour un modèle sans interactions. Considérons ensuite les modèles obtenus en ajoutant une interaction. Ceci va nous donner la série des modèles $A * B + C + D, A * C + B + D, A * D + B + C, A + B * C + D, A + B * D + C, A + B + C * D$. Nous avons utilisé la notation standard $A * B = A + B + A{:}B$, où $A{:}B$ représente les termes d'interaction d'ordre un entre A et B. Le modèle M_1 ayant un AIC inférieur est ainsi obtenu.

Remarquez que ajouter $A{:}B$ coûte un degré de liberté (car toutes les interactions entre A et B s'expriment en fonction de $(\alpha\beta)_{11}$), tandis que ajouter $A{:}D$ implique une perte de 2 degrés de liberté. En effet, dans ce dernier cas $A{:}D$ correspond avec 2 termes d'interactions indépendants : $(\alpha\delta)_{11}$ et $(\alpha\delta)_{12}$. Il découle des contraintes sur les paramètres d'interaction d'ordre 1 que $(\alpha\delta)_{21} = -(\alpha\delta)_{11}$, $(\alpha\delta)_{22} = -(\alpha\delta)_{21}$, $(\alpha\delta)_{23} = -(\alpha\delta)_{13}$, avec $(\alpha\delta)_{13} = -(\alpha\delta)_{12} - (\alpha\delta)_{11}$.

L'étape suivante consiste à supprimer de M_1 le terme réduisant le moins l'AIC. Un nouveau modèle M_2 sera ainsi examiné. Si le terme exclu est celui qui vient d'être ajouté à l'étape précédente, il sera conservé et un nouveau terme sera inclus. Et cette procédure continue jusqu'à ce qu'il n'y ait plus de diminution possible de l'AIC.

Finalement, nous désirons également avoir la possibilité de spécifier la propriété pour le modèle « optimal » de contenir un modèle minimal (« lower model ») et d'être un sous-modèle d'un modèle maximal (« upper model »).

Tableau 3.3 – Résultats de la procédure de sélection des variables

```
Stepwise Model Path Analysis of Deviance Table

Initial Model:  A + B + C + D

Final Model:    A + B + C + D +
    B:D + A:C + C:D + A:D + B:C + A:B + B:C:D

     Step Df  Deviance Resid. Df Resid. Dev      AIC
1                              18   153.8007 165.8007
2   + B:D -2 -45.71084         16   108.0899 124.0899
3   + A:C -1 -41.96315         15    66.1267  84.1267
4   + C:D -2 -21.52395         13    44.6028  66.6028
5   + A:D -2 -14.20011         11    30.4026  56.4026
6   + B:C -1  -6.32617         10    24.0765  52.0765
7 + B:C:D -2 -12.92004          8    11.1564  43.1564
8   + A:B -1  -4.40462          7     6.7518  40.7518
```

Dans notre exemple, le modèle $(A+B+C+D)$ qui est sans termes d'interaction est spécifié comme minimal, tandis que $(A+B+C+D)^3$ est le modèle maximal. Ce dernier contient les effets principaux et toutes les interactions jusqu'à l'ordre 2.

Le logiciel Splus permet de faire la sélection des variables en suivant cette méthodologie. Les ajustements se font avec les algorithmes utilisés pour l'estimation des Modèles Linéaires Généralisés. Rester dans le cadre des modèles linéaires généralisés nous donne l'avantage de pouvoir accéder à beaucoup d'outils déjà implémentés dans les logiciels statistiques. Les résultats de notre procédure de sélection se trouvent dans le tableau 3.3. Après 8 étapes, un modèle contenant toutes les interactions du premier ordre, et une seule interaction de deuxième ordre a été trouvé. Dans le tableau d'analyse de la déviance, la perte (ou le gain), en terme de degrés de liberté et la perte (ou le gain), en terme de déviance, sont présentées à chaque étape de l'algorithme.

Le modèle M, finalement choisi, a un nombre de degrés de liberté $df = 7$, et une déviance $G^2 = 6.75$. Par la formule (3.16), on trouve que $AIC = 40.7518$. Le modèle M s'écrit comme $(A+B+C+D)^2 + B{:}C{:}D$. Cette notation reste standard et homogène avec le paragraphe précédent. Nous avons, par exemple, $B{:}C{:}D = (\beta\gamma\delta)_{111} + (\beta\gamma\delta)_{112}$.

Estimons à présent le modèle M comme étant un modèle linéaire généralisé (MLG) du type de « Poisson ». Comme nous l'avons vu dans la section 3.2.4, ceci est une approche légitime. Les erreurs-types (SE) sont estimées avec des formules bien connues de la théorie des MLG. Des valeurs de $\hat{\theta}_j / SE(\hat{\theta}_j)$ supérieures,

Tableau 3.4 – Paramètres estimés, erreurs-types et valeurs de la statistique t pour le modèle $(A+B+C+D)^2+B:C:D$

	Value	Std. Error	t value
(Intercept)	3.59153427	0.03637291	98.742014
A	-0.27146023	0.03473699	-7.814732
B	-0.07035600	0.03537424	-1.988905
C	-0.12914513	0.03571661	-3.615828
D1	0.19151177	0.04616458	4.148457
D2	0.07083814	0.02385888	2.969047
A:B	-0.07289485	0.03477556	-2.096152
A:C	0.20744135	0.03425258	6.056225
AD1	0.11549493	0.04501262	2.565834
AD2	0.07123692	0.02360084	3.018406
B:C	-0.08646540	0.03483656	-2.482030
BD1	0.24997137	0.04373471	5.715628
BD2	-0.06709757	0.02333602	-2.875279
CD1	-0.09039252	0.04439964	-2.035884
CD2	0.07886837	0.02370421	3.327189
B:CD1	0.14023834	0.04347375	3.225816
B:CD2	0.04227349	0.02319477	1.822544

en valeur absolue, à 1.96, indiquent que les coefficients sont significatifs. Les résultats sont présentés dans le tableau 3.4. Nous remarquons ainsi qu'un grand nombre de coefficients sont proches de ceux estimés par le modèle $(A+B+C+D)^2$ et donnés dans le tableau 3.2. L'interprétation des valeurs numériques des coefficients dépendra de la façon dont la matrice des variables explicatives $\mathbf{X}$ a été construite. En effet, pour identifier les paramètres, différentes re-paramétrisations peuvent être effectuées. Ce problème est identique aux difficultés de choix des contrastes pour une analyse ANOVA.

Après avoir estimé le modèle sélectionné, il nous reste à le valider. Nos calculs nous indiquent que la déviance G^2 mène à une P-valeur de 0.45. Le modèle n'est donc pas contredit par les données. La dernière étape de la modélisation des données du tableau 3.1 consiste en l'analyse des résidus.

3.4.3 Analyse des résidus

L'analyse des résidus nous aidera à détecter d'éventuels points aberrants. Un tel point correspond à une cellule ne collant pas bien avec le modèle log-linéaire suivi par la très vaste majorité des données. Des valeurs des résidus importantes nous indiqueront l'existence de ces points. Le problème est maintenant de savoir à partir de quelle valeur un résidu peut être considéré comme étant « grand ». Les résidus bruts :

$$e_i = y_i - \hat{\mu}_i$$

ne sont évidemment pas interprétables. En effet, un effectif $y_i = 100$ ajusté par 110 a un résidu de 10. Il est cependant mieux représenté qu'un effectif $y_j = 5$ avec $\hat{\mu}_j = 14$ qui n'a un résidu que de 9. Une solution de ce problème est la standardisation. Supposons que les effectifs y_i viennent d'une distribution de Poisson de paramètre μ_i. Nous avons donc $E(Y_i) = \text{Var}(Y_i) = \mu_i$. Les *résidus de Pearson* seront ainsi :

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}. \tag{3.17}$$

Un graphique des résidus de Pearson en fonction de leur indice permettra de visualiser quels résidus sont supérieurs à 2 et donc susceptibles d'être aberrants. Le choix de cette valeur critique de 2 est motivé par l'approximation de la distribution des résidus standardisés par une normale centrée réduite. Afin de vérifier la validité de cet ajustement, un « qqplot » peut être examiné.

Une analyse des résidus de Pearson a été faite pour le modèle de notre exemple. Les figures 3.1(a) et 3.1(b), résumant les résultats, suggèrent qu'il n'y a pas de problème de points aberrants, mais indiquent également que l'approximation de la distribution des résidus par une normale n'est pas très satisfaisante.

Un problème de la standardisation de l'équation (3.17) est qu'elle ne tient pas compte du fait que les μ_i sont estimés. Il existe un résultat asymptotique qui permet d'avoir une approximation des résidus différente via une distribution normale :

$$y_i - \hat{\mu}_i \overset{d}{\approx} N(0, \hat{D}(I - \hat{A})),$$

où $\hat{D}$ est une matrice diagonale avec les $\hat{\mu}_i$ sur sa diagonale, et :

$$\hat{A} = \mathbf{X}(\mathbf{X}'\hat{D}\mathbf{X})^{-1}\mathbf{X}'\hat{D}. \tag{3.18}$$

Une meilleure standardisation des résidus bruts est alors donnée par *les résidus ajustés :*

$$r_i^{aj} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1 - \hat{a}_{ii})}}, \tag{3.19}$$

où $\hat{a}_{ii}$ est le i-ème élément sur la diagonale de la matrice A. La matrice $\mathbf{X}$ dans l'équation (3.18) est la matrice des variables explicatives.

Les figures 3.1(c) et 3.1(d) présentent les résidus ajustés ainsi que leur qqplot (toujours pour notre exemple). La dispersion des résidus est maintenant plus importante tandis que l'approximation normale n'est pas devenue meilleure. Remarquons qu'une forte déviation par rapport à la normalité pourrait mettre en cause les approximations effectuées, par un khi-carré, sur les statistiques de test.

La démarche adoptée ici est comparable à la « studentisation » des résidus en régression linéaire. En effet, ces derniers sont définis comme étant : $(y_i - \hat{y}_i)/(\hat{\sigma}\sqrt{1 - h_{ii}})$ où h_{ii} est le i-ème élément sur la diagonale de la matrice $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$, appelée « hat-matrix ». Il est bien connu que des quantités

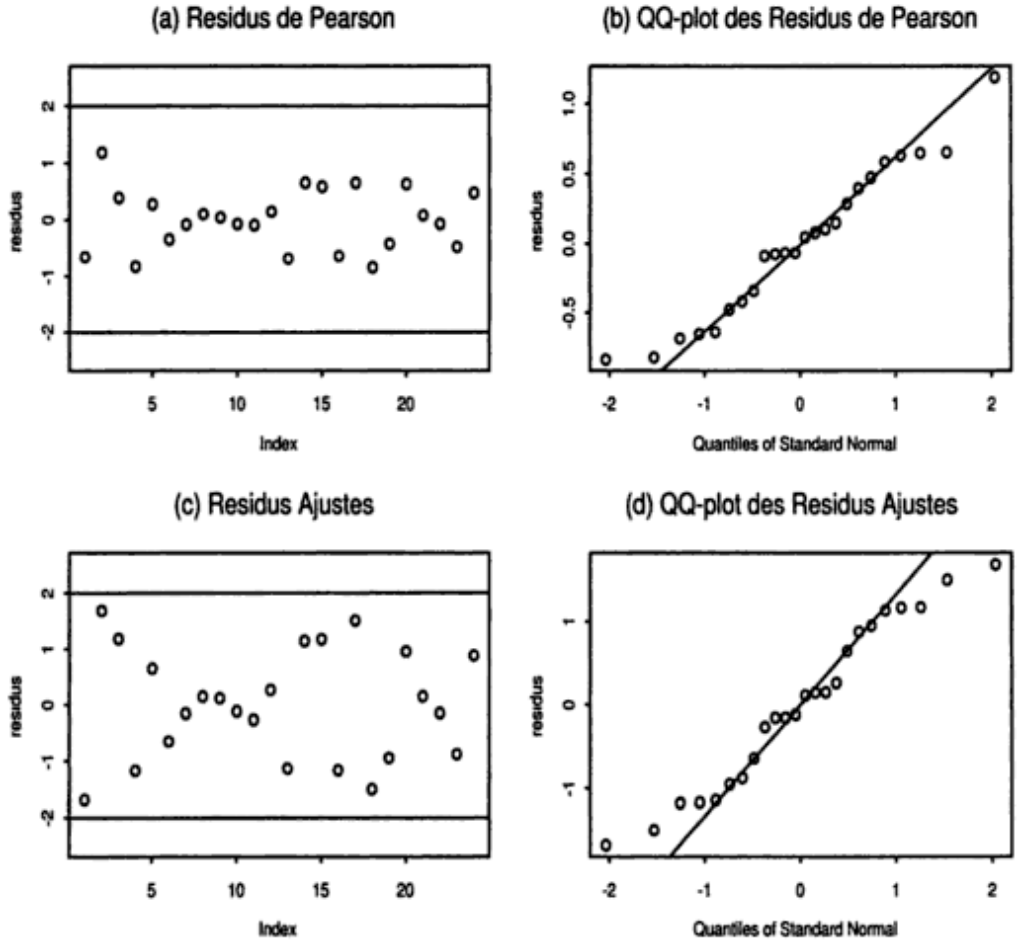

Figure 3.1 – Graphiques des résidus de Pearson et des résidus ajustés

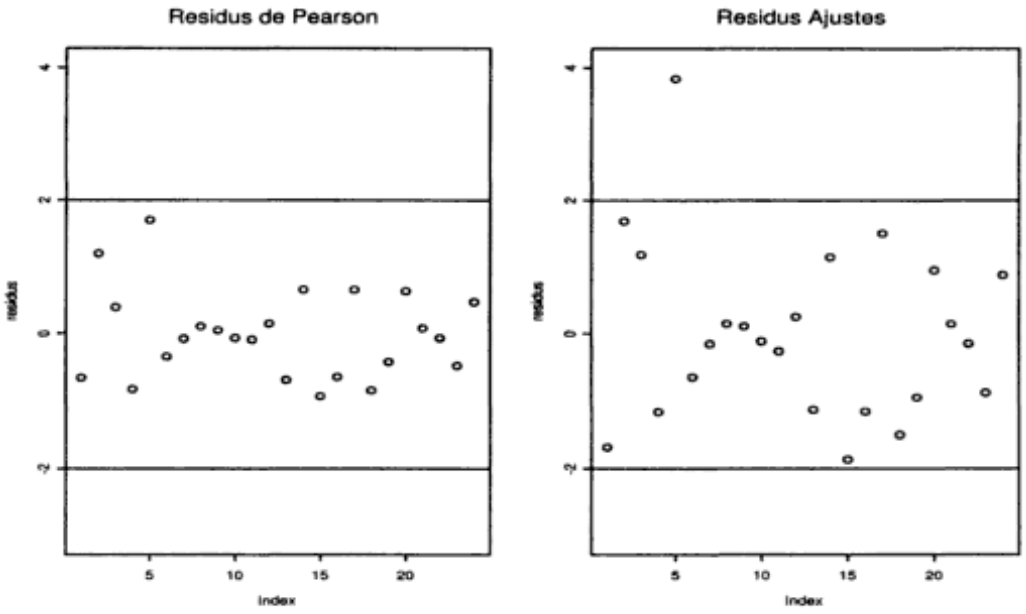

Figure 3.2 – Graphiques des résidus en présence de 2 points aberrants

h_{ii} élevées peuvent indiquer des points de levier lors de l'ajustement linéaire. Un point de levier a une influence disproportionnellement grande sur les estimations. Les $\hat{a}_{ii}$ jouent en principe un rôle comparable. Cependant la matrice $\mathbf{X}$ du modèle log-linéaire ne contient pas des réalisations de variables continues, mais uniquement des valeurs prises par des variables indicatrices. La présence de points de levier est donc peu probable, sauf peut-être pour des tables de contingence peu remplies (« sparse tables »). De toute façon, l'inspection des valeurs $\hat{a}_{ii}$ est toujours une bonne pratique.

Afin de vérifier si les diagnostics basés sur les résidus sont capables de détecter des points aberrants, nous avons augmenté la valeur de l'effectif numéro 5 dans le tableau 3.1 de 10, et diminué la valeur de l'effectif numéro 15, toujours de 10. La figure 3.2 représente le graphique des résidus. La cellule 5 est bien détectée comme « outlier » par les résidus ajustés, mais, en ce qui concerne la 15-ième, c'est nettement moins visible. Nous voyons tout de même qu'il est préférable de travailler avec les résidus ajustés plutôt qu'avec les résidus de Pearson. D'autres diagnostics, comme la distance de Cook, existent dans la littérature. Ils ne sont cependant pas performants en présence de points aberrants multiples. Des approches plus robustes sont nécessaires pour arriver à des diagnostics plus sûrs.

Pour conclure cette section, définissons encore les résidus de déviance par :

$$r_i^d = \text{sign}(y_i - \hat{\mu}_i)|2y_i \log \frac{y_i}{\hat{\mu}_i}|^{1/2} \tag{3.20}$$

pour $i = 1, \ldots, N$. Ces résidus sont toujours positifs, et la somme de leurs

carrés coïncide avec le critère de déviance. La contribution de l'observation i à la déviance est donnée par $(r_i^d)^2$, et doit rester petite.

3.5 Lien avec le modèle LOGIT

Le modèle log-linéaire étudie les associations entre plusieurs variables qualitatives. Si le nombre de variables est élevé, une multitude de paramètres seront à estimer, interpréter et considérer dans le processus de sélection du modèle. Déjà pour notre exemple, qui est simplement de dimension 4, les tableaux 3.2 et 3.4 sont loin d'être triviaux à lire et à interpréter. En pratique, le nombre de variables à incorporer dans le modèle doit rester restreint c'est-à-dire inférieur à 5 (selon Lebart et al., [1995, page 297]). Pour cette raison, l'analyse log-linéaire est parfois présentée comme une technique descriptive, aidant à explorer des tables de contingence.

Souvent une variable joue le rôle de variable dépendante et les autres d'explicatives. Dans des études de marketing, par exemple, on demande à un échantillon d'acheteurs potentiels, de répondre à une série de questions à choix multiples. Certaines caractéristiques propres aux personnes (sous forme de variables qualitatives) sont recueillies et finalement il leur est demandé si elles ont déjà acheté le produit examiné. Au lieu d' étudier toutes les associations possibles entre ces variables, il est envisageable de se contenter d'analyser l'association la plus intéressante, c'est-à-dire, la relation entre l'achat du produit (oui/non) et les caractéristiques intrinsèques des individus. Cette démarche est équivalente à celle qui consiste à entreprendre directement une analyse de type logit. Le modèle logit a dans ce cas l'avantage d'être plus simple à interpréter et il est donc préférable de l'utiliser ici.

Pour faciliter les notations, nous allons nous contenter d'étudier le lien entre le modèle log-linéaire et le modèle logit pour des tables de contingence de dimension 3.

3.5.1 Des tables de contingence de dimension 3 avec un facteur à expliquer

Désignons par X, Y et Z les variables qualitatives qui nous intéressent. La variable Y sera la variable à expliquer et les variables X et Z seront les explicatives. Le facteur X pourra prendre I modalités; Y et Z en prendront respectivement J et K. Le modèle saturé pour la table de contingence $I \times J \times K$ est donné par :

$$\log(\mu_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}, \quad (3.21)$$

pour $1 \leq i \leq I, 1 \leq j \leq J$ et $1 \leq k \leq K$. Prenons la dernière modalité J de Y comme niveau de référence. L'intérêt ici est d'étudier les chances d'avoir $Y = j$

par rapport à la catégorie de référence, conditionnellement aux valeurs prises par X et Z (plus précisément, la transformée logistique de ce « odds », donc le « log-odds »). Cela nous donne,

$$\log \frac{P(Y=j|X=i,Z=k)}{P(Y=J|X=i,Z=k)} = \log(\frac{\pi_{ijk}}{\pi_{iJk}}) = \log(\frac{\mu_{ijk}}{\mu_{iJk}}), \tag{3.22}$$

pour $j = 1, \ldots, J-1$. En substituant (3.21) dans l'équation (3.22), nous obtenons :

$$\begin{aligned} \log(\frac{\pi_{ijk}}{\pi_{iJk}}) &= [\beta_j - \beta_J] + [(\alpha\beta)_{ij} - (\alpha\beta)_{iJ}] + [(\beta\gamma)_{jk} - (\beta\gamma)_{Jk}] \\ &\quad + [(\alpha\beta\gamma)_{ijk} - (\alpha\beta\gamma)_{iJk}] \\ &= \lambda_j + \lambda^X_{j,i} + \lambda^Z_{j,k} + \lambda^{XZ}_{j,ik}, \end{aligned} \tag{3.23}$$

pour $j = 1, \ldots, J-1$. Les nouveaux paramètres $\lambda_j, \lambda^X_{j,i}, \lambda^Z_{j,k}$ et $\lambda^{XZ}_{j,ik}$ sont implicitement définis par l'équation ci-dessus. Les équations (3.23) sont appelées les équations logit. Les log-odds dans le membre de droite des équations logit s'écrivent en fonction des différents niveaux des variables X et Z, et en fonction des interactions entres les niveaux des facteurs explicatifs. Bien entendu, le modèle (3.23) est saturé et ses sous-modèles restent des modèles logit (généralisés). Par exemple, on posera souvent $\lambda^{XZ}_{j,ik} = 0$.

Le modèle caractérisé par l'équation (3.23) s'appelle un modèle logit généralisé. Dans le cas où $J = 2$, on obtient le modèle logit classique et l'équation (3.22) devient :

$$\text{logit}P(Y=1|X=i,Z=k) = \log(\frac{\pi_{i1k}}{1-\pi_{i1k}}) = \lambda + \lambda^X_i + \lambda^Z_k + \lambda^{XZ}_{ik}, \tag{3.24}$$

avec $\text{logit}(u) = \log(u/(1-u))$ la transformation logit. Nous pouvons remarquer qu'un modèle logit est un cas spécifique d'une régression logistique où toutes les variables explicatives sont qualitatives ou représentent des interactions entre des variables explicatives (donc comme un modèle ANOVA pour le logit d'une probabilité conditionnelle). Il est en effet possible de reparamétrer (3.24) comme $\text{logit}P(Y=1|U) = \boldsymbol{\theta}'u$ pour un certain vecteur u et un vecteur de paramètres inconnus $\boldsymbol{\theta}$. A son tour, le modèle de régression logistique est un exemple d'un modèle dichotomique, avec la fonction logit comme fonction de lien. De même, le modèle logit généralisé est un exemple d'un modèle polytomique.

3.5.2 Le modèle d'association homogène

Continuons avec les tableaux de dimension 3 et un facteur à expliquer à 2 modalités. Si nous excluons le terme d'interaction d'ordre 2 du modèle, nous obtenons ce que l'on appelle un modèle d'association homogène :

$$\log(\mu_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}. \tag{3.25}$$

Il suit que (3.24) se résume à un modèle logit sans interactions :

$$\text{logit}P(y=1|x=i,z=k)=\log(\frac{\pi_{i1k}}{1-\pi_{i1k}})=\lambda+\lambda_i^X+\lambda_k^Z \qquad (3.26)$$

Afin de motiver le choix du nom du modèle, prenons le cas ou $I = 2$. Calculons à l'aide de l'équation (3.24) le log-odds-ratio entre les facteurs X et Y, conditionnellement à Z (autrement dit, ce sont des log-odds ratios partiels). Celui-ci vaut :

$$\log\frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}=(\lambda+\lambda_1^X+\lambda_k^Z)-(\lambda+\lambda_2^X+\lambda_k^Z)=\lambda_1^X-\lambda_2^X$$

pour $k = 1,\ldots,K$. Le rapport des chances est devenu indépendant du niveau de Z. Un test de validité du modèle d'association homogène est donc équivalent à un test d'homogénéité des log-odds ratios. La comparaison du critère de déviance pour le modèle d'association homogène avec un percentile d'une $\chi^2(K-1)$ offre ainsi une alternative pour le test de Breslow-Day (cf. Chapitre 1). Ceci n'est pas réellement une surprise car, par de l'algèbre simple, il est possible de montrer que, pour le modèle saturé du tableau $2 \times 2 \times K$:

$$\log\frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}=4(\alpha\beta\gamma)_{11k}.$$

Cette équation permet de donner une interprétation aux interactions d'ordre 2 : ce sont en fait des log-odds-ratios partiels à une multiplication près. Gourieroux [1989, page 121] montre aussi que le terme d' interaction dans un modèle saturé pour un tableau 2×2 est égal à 4 fois le log-odds-ratio du tableau.

3.6 Modèles graphiques d'association

Un modèle graphique est un modèle log-linéaire qui est spécifié de façon unique par un graphe non-orienté. Les sommets de ce graphe représentent les variables, tandis que les arêtes correspondent à la présence d'une association partielle entre les 2 sommets/variables connectés. Deux variables n'ont pas d'association partielle, si elles sont indépendantes conditionnellement aux autres variables du modèle. Le graphe donne ainsi une représentation visuelle de l'équation du modèle. Contrairement à ce que fait, par exemple, l'analyse des correspondances, aucune visualisation de lignes et colonnes du tableau de contingence n'est représentée ici.

Le traitement théorique des modèles graphiques d'association nécessite l'introduction du langage des graphes et devient assez vite formel. Un livre de référence est celui de Whittaker [1990]. Une excellente introduction à ce sujet a été écrite par Fine [1992], à l'occasion des *Journées d'Etude en Statistique* organisées en 1990.

En ce qui nous concerne, nous allons essayer de montrer l'importance, pour l'utilisateur, des graphes d'association. Dans ce cadre de travail, la notion de graphe d'association sera considérée comme un synonyme de graphe d'indépendance conditionnelle.

3.6.1 Notations et définitions

Chaque modèle log-linéaire n'est en principe qu'une hypothèse nulle H_0, à tester par rapport à un modèle saturé. Cette H_0 peut être exprimée en termes matriciels, comme pour (3.7) ou en utilisant la notation de l'analyse de la variance, comme pour (3.15). Dans le cas où de nombreuses variables sont présentes, la notation introduite dans la section 3.4.2 s'avère plus intéressante. Notons A,B,C, ... les variables qualitatives qui nous concernent. Rappelons que :

- A représente les effets principaux associés aux niveaux du facteur A. On peut parler d' une interaction d'ordre 0.
- $A{:}B$ représente tous les termes d'interaction (d'ordre 1) entre les niveaux de A et ceux de B.
- $A * B = A + B + A{:}B$ représente tous les termes d'interactions d'ordre 1 et d'ordre 0 entre les variables A et B.
- $A{:}B{:}C$ représente tous les termes d'interaction (d'ordre 2) entre les niveaux de A, de B et de C.
- $A * B * C = A + B + C + A{:}B + A{:}C + B{:}C + A{:}B{:}C$ représente tous les termes d'interaction d'ordre 2, 1 ou 0, entre les variables A,B et C.
- ...

Limitons-nous ici aux modèles hiérarchiques (cf. section 3.4.2). En admettant une apparition multiple de certains termes, il sera toujours possible d'écrire l'équation du modèle en n'utilisant que les opérateurs + et $*$. Par exemple, $A*B*C+A*D+B*D+A{:}B{:}D = A*B*C+A*B*D$ ou $A*B+C+A{:}C = A*B+A*C$. Afin de simplifier la lecture, nous omettons le symbole * dans les multiplications, $A*B*C$ sera donc égal à ABC. Les termes des équations ainsi obtenues sont appelés les *marginales suffisantes* (cf. Andersen, [1997, page 88]).

L'équation du modèle détermine les indépendances conditionnelles entre les variables. Pour un sous-ensemble S de variables, $A \perp B \mid S$ indique que A et B sont indépendantes conditionnellement à S. Les graphes d'association permettent de détecter les associations entre variables. Pour conclure ce paragraphe, notons que $A \perp B$ désigne l'indépendance inconditionnelle entre les variables A et B.

3.6.2 Les graphes d'association

Un graphe d'association (ou d'indépendance conditionnelle) d'un modèle log-linéaire est un graphe non-orienté. Comme les sommets représentent les

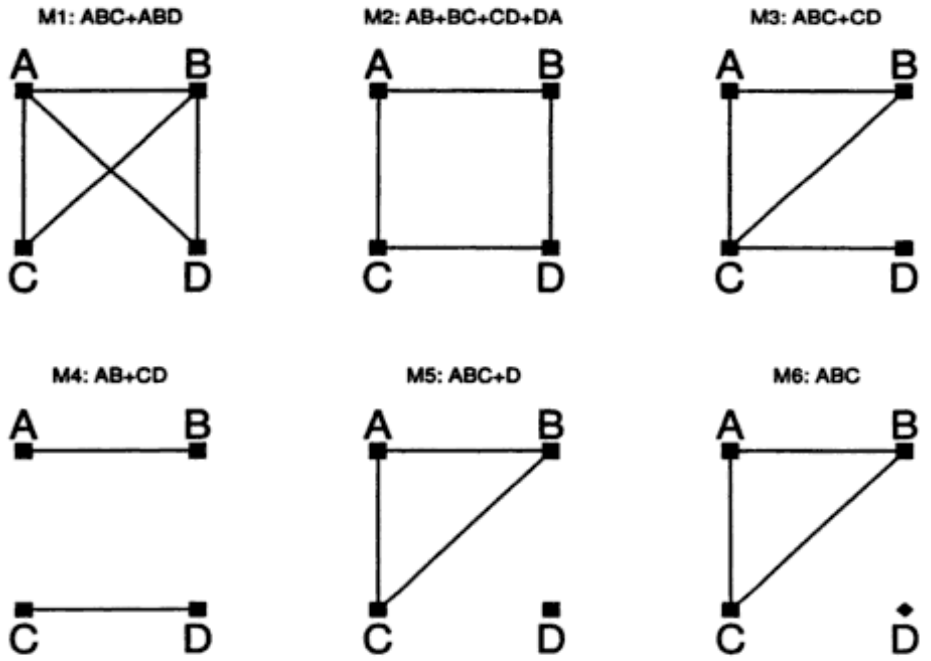

Figure 3.3 – Graphes d'associations pour 6 modèles différents

variables, il y aura autant de sommets que de facteurs dans le modèle. Les sommets reçoivent les mêmes étiquettes que les variables associées : $A, B, C, \ldots$. L'arête entre les sommets A et B sera tracée si et seulement s'il existe une dépendance conditionnelle (à toutes les autres variables du modèle) entre les 2 variables A et B. Ceci revient à dire que l'arête entre A et B figurera dans le graphe si et seulement s'il existe une marginale suffisante à laquelle A et B appartiennent de manière jointe. Donnons quelques exemples pour des modèles de dimension 4 :

$$M_1 : ABC + ABD \quad M_2 : AB + BC + CD + DA \quad M_3 : ABC + CD$$
$$M_4 : AB + CD \quad M_5 : ABC + D \quad M_6 : ABC.$$

Les graphes d'association de ces modèles se retrouvent dans la figure 3.3. En ce qui concerne la représentation de M_6, vu qu'il n'y a pas un effet principal pour le facteur D, le sommet qui s'y réfère est marqué par un losange. Il est possible de montrer que la distribution marginale de la quatrième variable D doit alors être uniforme. Pour le modèle $M_5 : ABC + D$, le sommet D n'a pas un marquage spécial car il existe un effet principal pour le facteur D. Une classification complète de tous les graphes d'association pour des tables de dimension 4 est présentée par Andersen [1997, page 93].

Commençons notre section explicative des graphes par un peu de syntaxe :

(a) Deux sommets A et B sont dits connectés s'il existe un chemin de A vers B.

(b) Si un sous-ensemble S des variables/sommets est tel que chaque chemin de A vers B passe par au moins un sommet de S, on dit que S sépare A et B.

Le théorème de séparation (cf. Whittaker [1990]) fournit 2 règles importantes :

1. $A \perp B \Leftrightarrow$ les sommets A et B ne sont pas connectés
2. $A \perp B \,|\, S \Leftrightarrow$ les sommets A et B sont séparés par S.

Ces règles se généralisent facilement au cas ou A et B sont des ensembles des variables/sommets.

A partir des graphes de la figure 3.3, nous voyons directement la structure de dépendance entre les variables. Pour M_1, C et D sont indépendants conditionnellement à A et B, donc $C \perp D \,|\, A, B$. Pour M_2, $A \perp D \,|\, B$, $A \perp D \,|\, C$, $B \perp C \,|\, A$ et $B \perp C \,|\, D$. Le modèle M_3 montre que conditionnellement à C, D est indépendant de A et B, tandis que M_4 impose l'indépendance entre {A,B} et {C,D}. Le modèle M_5 illustre l'indépendance de D par rapport à toutes les autres variables.

Il est possible que deux modèles différents aient le même graphe d'association. Par exemple, les modèles :

$$M_7 : AB + AC + BC + D \quad \text{et} \quad \mathrm{M}_8 : AB + AC + BC + CD$$

ont respectivement les même graphes d'association que M_5 et M_3. Ce type de représentation permet d'indiquer parfaitement les interactions d'ordre 1 entre les facteurs, mais risque de cacher l'absence des termes d'interaction d'ordre supérieur. Dans la section suivante, nous introduirons les *modèles graphiques d'association* qui jouissent de la propriété assurant que deux modèles graphiques d'association différents ne peuvent pas avoir le même graphe d'association.

3.6.3 Les modèles graphiques d'association

Notons par $\mathcal{G}(M)$ le graphe d'association correspondant à un modèle M. Un modèle graphique d'association $M_{\mathcal{G}}$ est défini comme étant un modèle hiérarchique satisfaisant la propriété stipulant que chaque modèle hiérarchique M avec $\mathcal{G}(M) = \mathcal{G}(M_{\mathcal{G}})$ doit être un sous-modèle de $M_{\mathcal{G}}$.

A partir d'un graphe $\mathcal{G}_0$ donné, il n'est pas difficile de trouver le modèle graphique d'association correspondant, en recherchant les *cliques* du graphe. Par clique nous entendons une collection de sommets respectant la propriété d'existence d'arête entre chaque paire de sommets. En plus, il n'existe pas de sommet, n'appartenant pas à la clique, lié avec tous les sommets de la clique. Les cliques constituent les marginales suffisantes du modèle $M_{\mathcal{G}_0}$, qui sera un modèle graphique d'association.

Prenons le graphe associé au modèle M_3 (cf. figure 3.3). Les cliques sont données par $\{A, B, C\}$ et $\{C, D\}$; M_3 est le modèle graphique associé. Le

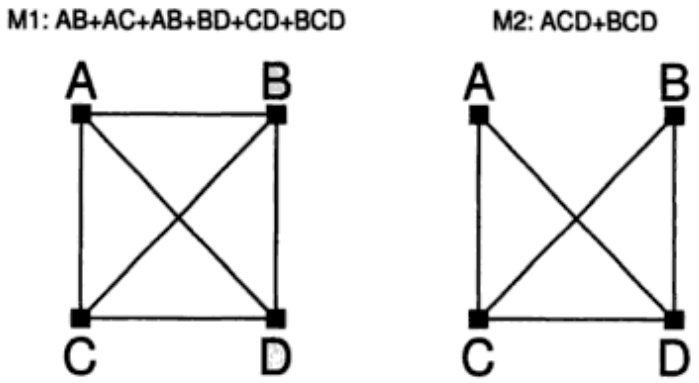

Figure 3.4 – Graphe d'association pour 2 modèles différents pour l'exemple des 1000 automobilistes

modèle M_8 a le même graphe que M_3 et est un sous-modèle de celui-ci ; M_8 n'est donc pas un modèle graphique. Contrairement aux modèles $M_1 - M_6$ qui sont des modèles graphiques, M_7 et M_8 n'en sont pas.

3.6.4 Choix du modèle graphique

Le choix d'un modèle graphique d'association se fait différemment du choix d'un modèle log-linéaire quelconque. Dans la section 3.4.2, nous avions proposé une méthode de minimisation du critère d'information d'Akaike, en recherchant un compromis entre simplicité du modèle et qualité d'ajustement. Ceci nous a souvent mené à des modèles sans termes d'interaction d'ordre élevé, car ceux-ci sont souvent peu significatifs (et difficilement interprétables). Reprenons notre exemple des 1000 automobilistes ; le modèle choisi était $M_1 : AB + AC + AD + BC + BD + CD + BCD$. Le graphe d'association qui s'y réfère est présenté dans la figure 3.4. Il n'est pas interprétable en tant que tel, car il est identique à celui d'un modèle saturé. D'autres procédures de sélection doivent être utilisées afin d'obtenir une représentation plus parlante.

Le but ici est de sélectionner un modèle d'association ayant une représentation graphique simple et ajustant de manière satisfaisante le tableau de contingence. Moins un graphe a d'arêtes, et donc moins de dépendances conditionnelles, plus il est facile à analyser. La procédure de sélection commence par le dessin d'un graphe complet (correspondant à un modèle saturé). Des nouveaux graphes sont crées en éliminant certaines arêtes. Chaque représentation ainsi obtenue correspond de manière unique avec un modèle graphique d'association. Plusieurs stratégies de sélection sont possibles (e.g. Edwards and Havranek [1985]). Une technique simple est la suivante :

1. Pour chaque graphe d'association nous éliminons une association partielle en supprimant une arête. Pour ce modèle restreint nous calculons le critère de déviance G^2, le nombre de degrés de liberté df, et le niveau de significativité observé (ou P-valeur) du test du rapport de maximum

de vraisemblance. En bref, nous calculons $P(\chi^2(\text{df}) > G^2)$. Nous faisons ceci pour chaque arête et donc pour chaque paire de variables. Tous ces niveaux de significativité observés peuvent être insérés dans une matrice $K \times K$, où K est le nombre de facteurs.

2. Nous considérons à ce moment le graphique obtenu en supprimant toutes les connections pour lesquelles les P-valeurs trouvées lors de la première étape sont supérieures à un α donné (par exemple $\alpha = 0.05$). Nous calculons maintenant le critère de déviance, le nombre de degrés de liberté et la P-valeur du nouveau modèle afin de vérifier si ce modèle ajuste bien le tableau de contingence.
3. Si le modèle obtenu lors de la seconde étape n'est pas satisfaisant, nous rajoutons séquentiellement des arêtes jusqu'au moment où nous obtenons un modèle non rejeté par le test du khi-carré. Les premières arêtes ajoutées sont évidemment celles qui procurent l'amélioration la plus importante. Si le modèle obtenu après l'étape 2 est par contre satisfaisant, nous pouvons soit le conserver en tant que tel, soit éventuellement essayer de le simplifier encore en enlevant d'autres arêtes.

Pour l'exemple des 1000 automobilistes, la matrice obtenue après la première étape est la suivante :

	A	B	C	D
A	*	0.093	0.000	0.006
B	0.093	*	0.003	0.000
C	0.000	0.003	*	0.000
D	0.006	0.000	0.000	*

En excluant uniquement AB, nous obtenons le graphe qui se trouve dans le panneau de droite de la figure 3.4. Le modèle graphique M_2 associé à celui-ci peut être obtenu très facilement puisque les cliques sont données par $\{A, C, D\}$ et $\{B, C, D\}$. L'équation du modèle devient alors $ACD+BCD$. A titre d'exemple, nous présentons également le cas où nous excluons l'arête AC. Nous obtenons ainsi le modèle graphique $ABD + BCD$, ce qui donne une P-valeur faible. En continuant, l'arête AB serait la seule qui se trouverait être « non-significative ». Le modèle résultant de la matrice ci-dessus est alors donné par $M_2 : ACD + BCD$. Le critère de déviance est $G^2 = 10.86$, les degrés de liberté $df = 6$, et le modèle n'est pas rejeté car la P-valeur est égale à 0.093 et donc supérieure à 0.05. (nous n'accepterions cependant pas ce modèle avec un $\alpha = 0.10$).

En conservant les notations et les règles de la section 3.6.2, l'interprétation du graphe du modèle M_2 (cf. figure 3.4) est $A \perp B \,|\, C, D$. Les variables A et B (usage professionnel de la voiture et respect de la vitesse maximale) se trouveraient être indépendantes, conditionnellement aux variables C et D (sexe et âge). Bien que A et B ne soient pas indépendantes (car liées par un chemin d'arêtes), elles pourraient êtres considérées comme telles, après avoir contrôlé pour le sexe et l'âge.

Tableau 3.5 – Fréquences observées pour toutes les combinaisons des facteurs A, B, C, D et E de l'exemple des congélateurs

A=Congélateur					
Oui	Non	B=Secteur	C=Revenu	D=Age	E=Sexe
152	39	Prive	Elevé	Âgé	Masculin
82	18	Public	Elevé	Âgé	Masculin
135	31	Prive	Moyen	Âgé	Masculin
35	12	Public	Moyen	Âgé	Masculin
89	45	Prive	Bas	Âgé	Masculin
20	9	Public	Bas	Âgé	Masculin
259	46	Prive	Elevé	Jeune	Masculin
101	26	Public	Elevé	Jeune	Masculin
183	55	Prive	Moyen	Jeune	Masculin
54	15	Public	Moyen	Jeune	Masculin
108	54	Prive	Bas	Jeune	Masculin
22	13	Public	Bas	Jeune	Masculin
82	17	Prive	Elevé	Âgé	Feminin
85	16	Public	Elevé	Âgé	Feminin
46	16	Prive	Moyen	Âgé	Feminin
60	11	Public	Moyen	Âgé	Feminin
29	29	Prive	Bas	Âgé	Feminin
40	18	Public	Bas	Âgé	Feminin
160	23	Prive	Elevé	Jeune	Feminin
152	28	Public	Elevé	Jeune	Feminin
89	17	Prive	Moyen	Jeune	Feminin
56	21	Public	Moyen	Jeune	Feminin
57	41	Prive	Bas	Jeune	Feminin
34	28	Public	Bas	Jeune	Feminin

3.6.5 Exemple

L'exemple suivant est tiré d'une étude danoise sur la prospérité (« Danish Welfare Study », cf. Andersen [1997]). L'échantillon contient 2758 individus qui ont été interviewés. Les 5 variables qualitatives suivantes ont été reprises : (A) : Ménage ayant un congélateur (Oui/Non), (B) : Secteur de l'emploi (Public/Privé), (C) : Revenu (Bas, Moyen, Elevé), (D) : Age (Jeune = moins de 40 ans, Âgé = les autres), (E) : Sexe (Masculin/Féminin). L'intérêt principal est porté sur la dépendance entre A et les 4 autres variables. Les données sont présentées dans le tableau 3.5.

Pour détecter les associations entre les variables, nous souhaitons proposer un modèle graphique d'association. La démarche adoptée sera celle décrite dans la section précédente. La première étape de la procédure est la création de la

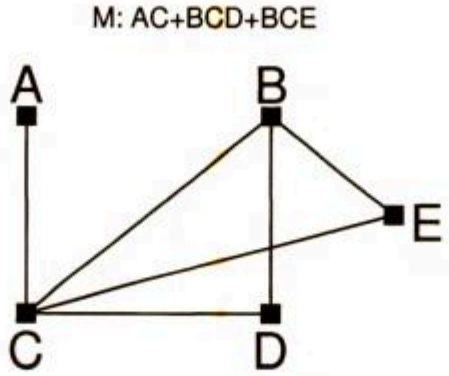

Figure 3.5 – Graphe d'association pour l'exemple des congélateurs

matrice des niveaux de significativité observés pour tous les modèles graphiques obtenus par l'élimination d'une arête du graphe complet. La création d'une telle matrice peut être faite automatiquement, en utilisant un logiciel statistique permettant la programmation. Les résultats obtenus sont :

	A	B	C	D	E
A	*	0.286	0.000	0.289	0.214
B	0.286	*	0.000	0.046	0.000
C	0.000	0.000	*	0.193	0.012
D	0.289	0.046	0.193	*	0.170
E	0.214	0.000	0.012	0.170	*

Nous allons essayer de nous débarrasser des arêtes A-B, A-D, A-E, C-D et D-E, qui correspondent à des valeurs P supérieures à 5%. Nous obtenons de cette manière le graphe de la figure 3.5. Les cliques de ce modèle sont $\{A, C\}, \{B, C, E\}$, et $\{B, D\}$. Le modèle graphique associé s'écrit $M_1 : AC + BCE + BD$. Il a une déviance de $G^2 = 40.35$ et des degrés de liberté $df = 31$. Le test du rapport de maximum de vraisemblance donne $P(\chi^2_{31} > 40.35) = 0.12$. Rien ne nous permet donc de rejeter le modèle M_1.

La conclusion la plus importante à tirer de cette analyse est que la variable A est indépendante de B, D et E, conditionnellement à C. Pour un revenu donné, le facteur de possession d'un congélateur est indépendant du secteur de l'emploi, de l'âge et du sexe du sujet. Autrement dit, le modèle détermine que le secteur de l'emploi, l'âge et le sexe n'influencent A que via l'intermédiaire du revenu auquel ils sont fortement associés.

Il est aussi remarquable que la variable D est indépendante de C et E, conditionnellement à B. Ceci est étonnant car on peut s'imaginer que le revenu (C) dépend de l'âge (D), conditionnellement au secteur de l'emploi (B). Il faut cependant être prudent avec l'interprétation afin d'éviter d'arriver à des conclusions hâtives. Tout d'abord, si l'échantillon était stratifié selon certains facteurs, des indépendances pourraient apparaître à cause du schéma d'échantillonnage. Ensuite, nous concluons que deux variables sont condition-

nellement indépendantes si l'élimination de l'arête correspondante ne rend pas le modèle incompatible avec les données. Ceci n'implique évidemment pas que ces variables soient réellement indépendantes mais elles sont modélisées comme étant indépendantes.

Remarque : « Collapsability »

Les tables de contingence de dimension 2 offrent quelques avantages importants : les distributions marginales sont faciles à représenter, il existe plusieurs techniques pour les analyser et les utilisateurs en ont une bonne connaissance pratique. Il peut être intéressant de réduire la dimension d'une table, en ne considérant que les distributions marginales deux à deux, afin de profiter de ces atouts. Avant de rentrer dans le vif du sujet, nous remarquons que dans ce paragraphe, les mots anglais « collaps » et « collapsability » seront traduits par réduire et« possibilité de réduction ».

Pour fixer les idées, supposons que nous nous intéressons à l'étude de l'association entre 2 variables qualitatives X et Y. Soit $S = \{z, v, w\}$ l'ensemble des variables que nous considérons en tant que variables de contrôle. Au lieu de travailler avec le tableau des effectifs $n_{i_1 i_2 i_3 i_4 i_5}$, nous pouvons nous limiter à l'étude du tableau réduit des fréquences marginales :

$$n_{i_1 i_2 \cdot \cdot} = \sum_{i_3, i_4, i_5} n_{i_1 i_2 i_3 i_4 i_5}.$$

La question naturelle à se poser ici est la suivante : est-ce que la liaison entre X et Y est la même en travaillant avec le tableau réduit plutôt qu'avec le tableau de dimension 5? En générale, la réponse est clairement négative. Il est intéressant de citer une propriété intéressante, connue sous le nom de « propriété de possibilité de réduction » :

> **Si $X \perp S \mid Y$ ou si $Y \perp S \mid X$, alors les paramètres d'association entre X et Y ne changent pas si l'on réduit le tableau en agissant sur les variables de S.**

Le graphe d'association est l'instrument parfait pour vérifier si la condition de possibilité de réduction est satisfaite. Reprenons notre exemple précédent. Nous déduisons de la figure 3.5 que $A \perp S \mid C$, avec $S = B, D, E$; la condition de possibilité de réduction est donc satisfaite. Le tableau réduit et le tableau des résidus ajustés correspondants sont donnés par :

		Revenu			
		Elevé	Moyen	Bas	Total
Congélateur	Oui	1073	658	399	2130
	Non	213	178	237	628
	Total	1286	836	636	2758