UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
Departamento de Estatística e Investigação Operacional

Estudo sobre a adaptação dos modelos log-lineares à ordinalidade e à presença de zeros amostrais em tabelas de contingência

Regina Maria Baltazar Bispo Carita

Mestrado em Probabilidades e Estatística

2007

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
Departamento de Estatística e Investigação Operacional

Estudo sobre a adaptação dos modelos log-lineares à ordinalidade e à presença de zeros amostrais em tabelas de contingência

Regina Maria Baltazar Bispo Carita

Dissertação orientada pelo Professor Doutor Dinis Pestana

Mestrado em Probabilidades e Estatística

2007

Ao Francisco e ao Pedro

Agradecimentos

Gostaria de agradecer a todos os que, directa ou indirectamente, contribuíram para a realização deste mestrado. Em particular, agradeço

- Ao Professor Doutor Dinis Pestana, pela orientação, pelas sugestões e por ser uma constante inspiração. Agradeço também as reuniões de trabalho sempre bem dispostas;
- À Dra. Ana Melo que recolheu e amavelmente cedeu os dados para análise;
- À minha colega Luísa Cunha por me acompanhar nesta caminhada e pela amiga que tem sido ao longo dos anos;
- Aos meus colegas do mestrado, Zé e Rui, pela constante boa disposição e camaradagem e por tornarem uma satisfação diária reencontrá-los;
- Quase em último, mas sempre em primeiro, ao João Nuno, pelos fins-de-semana, pelas birras aturadas, pelos mimos, pela calma, pelo amor e pela paciência do tamanho do mundo que nunca se esgota. Se não fosse ele nunca teria conseguido;
- Aos meus filhotes, Francisco e Pedro, porque é por eles e a pensar neles que tudo faz sentido;
- Aos meus pais e irmã que sempre me incentivaram a ir mais além;
- Aos meus sogros e cunhados por me ajudarem e compreenderem os fins-de-semana ausentes. Sem a sua ajuda teria sido impossível escrever esta tese.

Resumo

Os modelos log-lineares são, desde há muito, uma das mais importantes ferramentas na análise de tabelas de contingência, tendo impacto em muitas áreas científicas. Com a proliferação de *packages* informáticos capazes de executar os cálculos necessários e apresentar os resultados desejados, a utilização destes modelos tem crescido grandemente. Contudo, a análise de bases de dados de dimensões cada vez maiores, tem aumentado o grau de complexidade e de atipicidade da informação recolhida, tornando necessário o desenvolvimento de novas técnicas e a adaptação dos procedimentos clássicos a problemas específicos.

Neste trabalho aborda-se a problemática da adaptação dos modelos log-lineares clássicos à ordinalidade das variáveis e à existência de zeros em tabelas de contingência. Neste sentido, são abordados os modelos log-lineares ordinais que permitem descrever padrões de associação e interacção (ou a sua ausência) inerentes à ordinalidade das variáveis. Por outro lado, aborda-se o problema da existência de celas com zeros amostras, na perspectiva de construção de modelos quase-log-lineares. A abordagem das metodologias é estendida ao caso de coexistência de escalas ordinais e presença de zeros, abordando-se os modelos quase-log-lineares ordinais.

As metodologias abordadas foram aplicadas a um conjunto de dados enfatizando a sua aplicabilidade na análise de dados reais. Os dados, relativos a escalas nutricionais de crianças e organizados numa tabela tridimensional, permitiram, numa perspectiva de intervenção precoce e preventiva da obesidade infantil, estimar a probabilidade de uma criança ser obesa, ou possuir excesso de peso, num determinado contexto nutricional familiar.

Palavras chave: Modelos log-lineares; Variáveis ordinais; Zeros Amostrais; Zeros Estruturais; Tabelas de contingência.

Abstract

Log-linear models are, since a long time, one of the most used methodologies in contingency tables analysis, with impact on many scientific fields. With the growing availability of statistical packages that perform the necessary calculations and present the so desired results, these models have been increasingly used. Yet the analysis of bigger amounts of data have been incrementing the complexity and the messiness of collected data, implying the development of new technics and the adaptation of classical procedures to specific problems.

This study deals with the adaptation of classical log-linear models to ordinal cross classified data with zero counts. These models are adapted to the ordinality of the variables by including association and interaction terms that reflect the hierarchical characteristics of the categorical variables. Quasi-log-linear models are fitted to overcome estimated random zeros.

Studied methods were then applied to real data emphasizing the applicability of this methodology in data analysis. Parent and offsprings body structure data organized into a three-way contingency table, allow to determine the probability of a child being obese (or having excess of weight) in a particular nutritional family context.

Key words: Log-linear models; Ordinal variables; Sampling zeros; Structural zeros; Contingency tables.

Conteúdo

Capítulo 1

Introdução

As variáveis qualitativas são frequentes em muitos domínios científicos. A classificação de um conjunto de indivíduos em função de variáveis qualitativas leva à contagem do número de casos afectos às categorias. A análise e modelação deste tipo de dados - contagens - é, desde há muito, um desafio para estatísticos e investigadores que nos seus estudos lidam com este tipo de informação.

Com o avanço da tecnologia e o aumento da capacidade de análise de bases de dados de dimensões cada vez maiores, tem crescido o grau de complexidade e de atipicidade da informação recolhida[1]. A necessidade de novas técnicas e da adaptação de procedimentos padronizados tem-se, por isso, tornado uma realidade.

São muitas as investigações conduzidas nesta área do conhecimento que procuram dar resposta a problemas específicos relacionados com a análise de tabelas de contingência. Neste trabalho procurou-se ir ao encontro de preocupações frequentemente expressas por quem lida com este tipo de dados. Por um lado, o tratamento de escalas ordinais e, por outro, a existência de zeros em tabelas de contingência.

No que respeita à presença de variáveis ordinais, na grande maioria das situações são usados métodos "standard" para variáveis nominais, "desperdiçando" informação importante contida na relação de ordem dos dados. Por outro lado, a presença de zeros em tabelas de contingência é para muitos investigadores um dos maiores problemas aquando da modelação dos dados. Frequentemente é descrita a aplicação de "remédios" como sejam a combinação de categorias de variáveis de modo a aumentar a frequência das celas, o preenchimento das celas com algum tipo de "valores apropriados" , ou, a desmontagem dos dados até ao "desaparecimento" dos zeros, sendo frequentemente relatadas conclusões confusas ou inapropriadas. Dados deste tipo podem mesmo fazer com que os investigadores abandonem os respectivos estudos (Bishop, Fienberg and Holland, 1980).

[1] Nonstandard sets of data; Messy data

Com este trabalho pretendeu-se:

1. Abordar metodologias que permitam ultrapassar as dificuldades de modelação em tabelas de contingência com escalas ordinais e células nulas. São, por isso, abordados: *(a)* os métodos ordinais que usam a informação relativa à hierarquia das categorias, sendo exposta a vantagem que dai advém nomeadamente no que respeita à simplificação de modelos, e *(b)* as metodologias que permitem ultrapassar o problema da existência de celas com frequência nula, dando-se ênfase ao problema da estimação, selecção do melhor modelo e cálculo de graus de liberdade.

   A abordagem das metodologias é estendida ao caso de coexistência de escalas ordinais e presença de zeros.

2. Aplicar as metodologias abordadas a um conjunto de dados de forma a:

   - Enfatizar a aplicabilidade das metodologias expostas na análise de dados reais;
   - Analisar dados referentes a escalas nutricionais de crianças, contribuindo para, numa perspectiva de intervenção precoce e preventiva da obesidade infantil, estimar a probabilidade de uma criança ser obesa ou possuir excesso de peso num determinado contexto nutricional familiar.

De salientar que neste texto a análise é restringida ao caso de tabelas bi- e tridimensionais e em que ambas as variáveis são ordinais, embora os métodos apresentados sejam generalizáveis a tabelas de maior dimensão e aos casos em nem todas as variáveis sejam ordinais.

Esta dissertação está organizada em 9 capítulos. Nos capítulos 1 a 5 abordam-se os conceitos prévios (capítulo 1), os modelos probabilísticos (capítulo 2) e os modelos log-lineares clássicos (capítulos 3 a 5) que estão na base da adaptação das metodologias à presença de escalas ordinais (capítulo 6) e à existência de celas com frequência nula (capítulo 7). No capítulo 8 serão apresentados os resultados da análise dos dados reais e no capítulo 9, apresenta-se a conclusão do trabalho.

Capítulo 2

Conceitos prévios

2.1 Variáveis qualitativas

São designadas por variáveis qualitativas aquelas cujas modalidades são definidas por categorias[1] exaustivas e mutuamente exclusivas. A classificação de um conjunto de indivíduos em função de variáveis qualitativas possibilita a contagem do número de casos afectos a cada uma das categorias. Os modelos estatísticos abordados no presente trabalho — Modelos Log-lineares — adequam-se precisamente à análise deste tipo de dados: contagens ou frequências.

As variáveis qualitativas podem subdividir-se em dois grandes grupos. Quando entre as categorias não é possível estabelecer uma relação de ordem, as variáveis são classificadas como nominais. Ao contrário, quando as categorias são hierarquizáveis, as variáveis dizem-se ordinais. As escalas ordinais são frequentemente usadas, *e.g.*, nas Ciências Sociais, Biomédicas, Psicológicas e da Educação. Em particular, em Medicina, são muitas as variáveis ordinais referidas na bibliografia deste domínio científico. A título de exemplo, mencionam-se a Escala de Apgar, as Escalas de Avaliação Neurológica e as Escalas de Classificação Nutricional.

2.2 Tabelas de contingência

As *tabelas de contingência* não são mais que uma forma de apresentação tabelar de contagens de efectivos de classes (Pestana, 2006). Uma tabela com l linhas e c colunas, diz-se que tem *dimensão* $l \times c$ sendo, por isso, referenciada como uma tabela $l \times c$ e designada por *bidimensional*. Da mesma forma, uma tabela com l linhas, c colunas e s estratos, é referenciada como uma tabela $l \times c \times s$ sendo designada por *tridimensional*. Para dimensões superiores as tabelas são designadas por *multidimensionais*.

[1] Também designadas na literatura por classes.

Considere-se a situação em que se possui uma amostra aleatória de dimensão n sobre a qual se tem informação de duas variáveis categorizadas, X com l categorias e Y com c categorias. Os dados relativos às contagens do número de casos observados no cruzamento das categorias das duas variáveis podem ser organizados numa tabela com l linhas, representando as categorias da variável X, e c colunas, representando as categorias da variável Y. As celas da tabela representam os lc possíveis resultados correspondentes às combinações das categorias. Nesta situação está-se perante uma *tabela de margens livres* (Pestana, 2006). De uma forma genérica, uma tabela de contingência bidimensional resultante de uma dupla classificação pode ser apresentada tal como consta na Tabela 2.1.

Tabela 2.1: Tabela de contingência bidimensional

	Variável Y					
Variável X	1	$\cdots$	j	$\cdots$	c	*Total*
1	o_{11}	$\cdots$	o_{1j}	$\cdots$	o_{1c}	$o_{1\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	o_{i1}	$\cdots$	o_{ij}	$\cdots$	o_{ic}	$o_{i\bullet}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
l	o_{l1}	$\cdots$	o_{lj}	$\cdots$	o_{lc}	$o_{l\bullet}$
Total	$o_{\bullet 1}$	$\cdots$	$o_{\bullet j}$	$\cdots$	$o_{\bullet c}$	n

Há, no entanto, outro tipo de tabelas. Quando várias amostras são classificadas segundo um determinado critério, ou seja, c amostras são classificadas em função das l categorias de uma variável X, está-se perante *tabelas com uma margem fixa e uma margem livre*. Finalmente, há também situações onde os esquemas amostrais conduzem a *tabelas com ambas as margens fixas* (Pestana, 2006).

2.2.1 Notação

A matriz de frequências observadas numa tabela bidimensional pode representar-se por $\{o_{ij}\}$, $(i = 1, ..., l; j = 1, ..., c)$ tal que

$$\{o_{ij}\} = \begin{pmatrix} o_{11} & \cdots & o_{1c} \\ \vdots & \ddots & \vdots \\ o_{l1} & \cdots & o_{lc} \end{pmatrix}$$

As frequências marginais são então representadas por $o_{\bullet j}$ e $o_{i\bullet}$ onde, convencionalmente, o ponto substitui o índice relativamente ao qual se efectuou a soma

$$o_{\bullet j} = \sum_{i=1}^{l} o_{ij} \qquad o_{i\bullet} = \sum_{j=1}^{c} o_{ij}$$

O número total de observações (n) satisfaz as proposições

$$n = \sum_{i=1}^{l} o_{i\bullet} = \sum_{j=1}^{c} o_{\bullet j} = \sum_{i=1}^{l} \sum_{j=1}^{c} o_{ij}$$

Dada a matriz de frequências observadas $\{o_{ij}\}$, representar-se-à por p_{ij} a proporção de casos classificados na cela (i, j), isto é, $p_{ij} = o_{ij}/n$, sendo, naturalmente, $\sum_i \sum_j p_{ij} = 1$. A matriz $\{p_{ij}\}$ representa a *Distribuição Empírica Conjunta*. As *Distribuições Empíricas Marginais* são representadas pelos vectores dos totais marginais $\{p_{i\bullet}\}$ e $\{p_{\bullet j}\}$.

Numa tabela de contingência pode ainda definir-se a proporção de casos classificados na categoria i de X, dado que pertencem à classe j de Y, isto é, a proporção condicionada $p_{i(j)} = o_{ij}/o_{\bullet j}$.

A notação referida até agora refere-se apenas à amostra. Na população, p é substituída pela letra Grega π, isto é, $\{\pi_{ij}\}$ representará as probabilidades conjuntas, $\{\pi_{i\bullet}\}$ e $\{\pi_{\bullet j}\}$ as probabilidades marginais e $\{\pi_{i(j)}\}$ as probabilidades condicionadas.

Naturalmente, em tabelas multidimensionais, e em particular em tabelas tridimensionais, usar-se-à notação similar, aumentando o número de índices em função do número de variáveis envolvidas.

2.3 Razão de produtos cruzados

Um dos objectivos principais da análise de tabelas de contingência é determinar se as variáveis envolvidas na classificação dos casos são ou não independentes. Adicionalmente, no caso da não independência, existe a necessidade de quantificar o grau de associação das variáveis. Para este efeito, existem várias medidas de associação extensamente abordadas na literatura (*e.g.* Upton, 1980, capítulos 2, 3 e 4). No âmbito dos modelos log-lineares, tem especial interesse a estatística *Razão de Produtos Cruzados*, também designada por *Razão de Chances*.

Considerando a linha 1 de uma tabela 2×2, a *chance* de obter uma resposta na coluna 2 em vez de a obter na coluna 1 é dada por (Agresti, 1984, página 15)

$$\Omega_1 = \frac{\pi_{12}}{\pi_{11}}$$

Da mesma forma, na linha 2, a *chance* de obter uma resposta na coluna 2 em vez de a obter na coluna 1 é dada por

$$\Omega_2 = \frac{\pi_{22}}{\pi_{21}}$$

Uma *chance* igual a 1 significa uma situação de equiprobabilidade de classificação em qualquer uma das colunas; um valor superior a 1 mostrará que a probabilidade de classificação na coluna 2 é maior que a probabilidade de classificação na coluna 1.

A razão entre Ω_1 e Ω_2

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{2.1}$$

define a *Razão de Produtos Cruzados* ou a *Razão de Chances*. Esta estatística, estimada por $\hat{\theta} = (o_{11}o_{22})/(o_{12}o_{21})$, mede o grau de associação entre as duas variáveis. Para probabilidades não nulas, o seu valor será sempre não negativo e igual a 1 quando as variáveis forem independentes. Um valor inferior a 1 indicará que os elementos da linha 1 apresentam uma menor chance de serem classificados na coluna 1 que os elementos da linha 2 e, inversamente, um valor superior a 1 indicará que os elementos da linha 1 apresentam uma maior chance de serem classificados na coluna 1 que os elementos da linha 2.

Em tabelas $l \times c$, a razão de produtos cruzados pode calcular-se combinando os C_2^l pares de linhas com os C_2^c pares de colunas, isto é, existem $C_2^l \times C_2^c$ Razões de Chances numa tabela bidimensional com l linhas e c colunas. Contudo, neste conjunto de estatísticas existe muita informação redundante. É possível mostrar que a família de razões calculadas mediante o uso de linhas adjacentes e colunas adjacentes, designadas por *Razão de Chances Locais* e definidas por

$$\theta_{ij} = \frac{\pi_{i,j}\ \pi_{i+1,j+1}}{\pi_{i+1,j}\ \pi_{i,j+1}} \quad i = 1, ..., l-1;\ j = 1, ..., c-1 \tag{2.2}$$

determinam todas as $C_2^l \times C_2^c$ Razões de Chances numa tabela $l \times c$ (Agresti, 1990, página 18). O valor desta estatística descreve a magnitude relativa das associações "locais" na tabela.

Em tabelas tridimensionais, definem-se as *Razões de Produtos Cruzados Condicionais* que avaliam a associação condicional local entre duas variáveis, dado um certo nível da terceira variável. Assim, por exemplo, a Razão de Chances entre X e Y, dado o nível k de Z, $\theta_{ij(k)}$, define-se como

$$\theta_{ij(k)} = \frac{\pi_{ijk}\ \pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\ \pi_{i+1,j,k}} \quad i = 1, ..., l-1;\ j = 1, ..., c-1 \tag{2.3}$$

As Razões de Chances $\theta_{i(j)k}$ e $\theta_{(i)jk}$ são, naturalmente, definidas de modo análogo. Adicionalmente, o rácio entre as Razões de Produtos Cruzados Condicionais

$$\theta_{ijk} = \frac{\theta_{ij(k+1)}}{\theta_{ij(k)}} = \frac{\theta_{i(j+1)k}}{\theta_{i(j)k}} = \frac{\theta_{(i+1)jk}}{\theta_{(i)jk}} \tag{2.4}$$

permite descrever a interacção de 2ª ordem local, isto é, θ_{ijk} representa a interacção numa secção de dimensão $2 \times 2 \times 2$ numa tabela tridimensional constituída por linhas, colunas e estratos adjacentes. Na ausência de interacção de 2ª ordem, $\theta_{ijk} = 1$ (Agresti, 1990, página 278).

Capítulo 3

Modelos probabilísticos

Numa tabela de contingência existem três modelos probabilísticos adequados para explicar o número de ocorrências dependendo do esquema amostral adoptado e dos objectivos da análise (Paulino e Singer, 2006, capítulo 2):

1. n **aleatório**:
   As frequências numa tabela de contingência por serem contagens de efectivos de classes assumem valores inteiros não negativos. Considerando o número total de observações à partida indeterminado, sendo realizadas tantas observações quanto possível num dado intervalo de tempo[1], então as frequências observadas numa tabela bidimensional o_{ij} $(i = 1, ..., l; j = 1, ..., c)$ não são mais que realizações independentes das variáveis aleatórias O_{ij} bem modeladas por uma Distribuição Poisson com valor esperado $E(O_{ij}) = e_{ij}$ $(e_{ij} > 0$, com $i = 1, ..., l$ e $j = 1, ..., c)$

$$\{O_{ij}\} \frown Poisson(\{e_{ij}\}) \tag{3.1}$$

   Admitindo O_{ij} $(i = 1, ..., l; j = 1, ..., c)$ independentes e identicamente distribuídas, a função massa de probabilidade conjunta é dada por

$$f(\{o_{ij}\}|\ \{e_{ij}\}) = \prod_{i=1}^{l}\prod_{j=1}^{c} f(o_{ij}|\ e_{ij}) = \prod_{i=1}^{l}\prod_{j=1}^{c} \frac{\exp(-e_{ij})\ e_{ij}^{o_{ij}}}{o_{ij}!} \tag{3.2}$$

   com $o_{ij} \in \mathbb{N}_0$ e $e_{ij} \in \mathbb{R}^+$.

   Numa tabela tridimensional $l \times c \times s$ ter-se-à

[1] O modelo baseado em Processos de Poisson pode não ser adequado para dados definidos espacialmente pois a existência, por exemplo, de correlação entre regiões contíguas não é compatível com a hipótese de independência entre as contagens por área (Paulino e Singer, 2006, página 25).

$$f(\{o_{ijk}\}|\ \{e_{ijk}\}) = \prod_{i=1}^{l}\prod_{j=1}^{c}\prod_{k=1}^{s}\frac{\exp(-e_{ijk})\ e_{ijk}^{o_{ijk}}}{o_{ijk}!} \tag{3.3}$$

com $o_{ijk} \in \mathbb{N}_0$ e $e_{ijk} \in \mathbb{R}^+$ $(i = 1, ..., l; j = 1, ..., c; k = 1, ..., s)$.

2. n **fixo**:

   Um segundo esquema amostral, na prática mais frequente, consiste em fixar n. Considerando o caso bidimensional, dado que $n = \sum_i \sum_j o_{ij}$, então, naturalmente, o_{ij} nunca poderá exceder n. Nestas condições as variáveis O_{ij} são dependentes. O vector de frequências observadas O_{ij} $(i = 1, ..., l; j = 1, ..., c)$ condicionadas por n, isto é, pela sua soma, têm Distribuição Multinomial

$$\{O_{ij}\}|\ n \frown Multinomial(n, \{\pi_{ij}\}) \tag{3.4}$$

   com valor esperado $e_{ij} = n\pi_{ij}$, isto é, $\pi_{ij} = \dfrac{e_{ij}}{n} = \dfrac{e_{ij}}{\sum_i \sum_j e_{ij}}$.

   e função massa de probabilidade

$$f(\{o_{ij}\}|\ n, \{\pi_{ij}\}) = \frac{n!}{\prod_i \prod_j o_{ij}} \prod_i \prod_j \pi_{ij}^{o_{ij}} \tag{3.5}$$

   Generalizando ao caso de tabelas tridimensionais, tem-se:

$$f(\{o_{ijk}\}|\ n, \{\pi_{ijk}\}) = \frac{n!}{\prod_i \prod_j \prod_k o_{ijk}!} \prod_i \prod_j \prod_k \pi_{ijk}^{o_{ijk}} \tag{3.6}$$

3. **Totais marginais fixos**:

   Uma alternativa aos esquemas amostrais anteriormente expostos consiste em fixar antecipadamente o número de elementos pertencentes a várias amostras independentes correspondentes, por exemplo, às c categorias da variável Y. Este delineamento corresponde a um esquema de amostragem estratificada, no qual, se extrai independentemente uma amostra aleatória simples de cada um dos estratos considerados. Ou seja, trata-se de um esquema amostral conducente à fixação de uma das margens da tabela. Neste caso, considerando o caso bidimensional, a probabilidade de classificação do elemento w $(w = 1, ..., n)$ na cela (i, j) é dada pela a probabilidade do elemento ser classificado na categoria i de X, dado que pertence à classe j de Y, isto é, trata-se da probabilidade condicionada

$$\pi_{i(j)} = P(X_w = i|\ Y_w = j)(i = 1, ..., l; j = 1, ..., c)$$

Assim, considerando fixos os totais marginais, as contagens relativas a cada uma das l linhas de X têm distribuições binomiais de parâmetros $o_{i\bullet}$ e $\pi_{i(j)}$ $(i = 1, ..., l; j = 1, ..., c)$, isto é, distribuições multinomiais univariadas independentes. O modelo probabilístico adequado à modelação de O_{ij} é portanto o Modelo Produto de Multinomiais cuja função massa de probabilidade é dada por

$$f(\{o_{ij}\}|\ \{o_{i\bullet}\}, \{\pi_{i(j)}\}) = \prod_{i=1}^{l} \left(\frac{o_{i\bullet}!}{\prod_{j=1}^{c} o_{ij}!} \prod_{j=1}^{c} \pi_{i(j)}^{o_{ij}} \right) \tag{3.7}$$

Os modelos probabilísticos expostos, apesar de decorrerem de diferentes esquemas amostrais, estão intimamente associados entre si (Paulino e Singer, 2006, página 26). Na realidade, é possível mostrar que o *kernel*[2] da função verosimilhança é idêntico para os três esquemas amostrais expostos (Bishop, Fienberg and Holland, 1980, página 64). Daqui decorre que os estimadores de máxima verosimilhança para as frequências esperadas das celas de uma tabela de contingência sob os três esquemas de amostragem abordados são os mesmos (Bishop, Fienberg and Holland, 1980, página 448). Assim, a análise inferencial de uma tabela de contingência sob amostragem de Poisson é equivalente à análise sob amostragem multinomial ou produto-multinomial.

[2]Neste contexto, entende-se por *kernel* a parte da função de probabilidade que depende dos parâmetros.

Capítulo 4

Modelos log-lineares

A análise clássica de tabelas de contingência é feita mediante análises bivariadas, através do cálculo de estatísticas que permitem detectar e medir o grau de associação entre duas variáveis. Alternativamente, o ajustamento de modelos log-lineares permite analisar associações simples ou complexas entre duas ou mais variáveis. Este tipo de abordagem possibilita uma abordagem sistematizada de tabelas multidimensionais e permite estimar a magnitude os efeitos de interesse e, consequentemente, estabelecer uma hierarquia entre as variáveis no que respeita à sua importância relativa (Everitt, 1994, páginas 73-74).

A expressão *modelo* refere-se a uma estrutura conceptual ajustável às observações. A sua construção implica a definição dos termos que o compõem e a estimação dos parâmetros. O número de termos depende da dimensionalidade da tabela. O número de parâmetros de cada termo depende do número de categorias de cada variável. As estimativas dos parâmetros do modelo fornecem informação sobre a magnitude do efeito que as variáveis, ou combinações de variáveis, têm na determinação dos valores observados.

A maior facilidade no tratamento matemático de estruturas matemáticas lineares do que no tratamento de estruturas não lineares, leva, em muitas circunstâncias, à linearização dos modelos estruturais. Na análise de padrões de associação em tabelas de contingência, a logaritmização das probabilidades das celas de tabelas de contingência permite essa linearização e dá, assim, origem aos designados *Modelos Log-lineares*. Um modelo com um número de parâmetros igual ao número de celas da tabela é designado por *Saturado*. Estes modelos possuem um ajustamento perfeito já que, nesta situação, as frequências esperadas são, simplesmente, as próprias frequências observadas. A eliminação de termos do modelo log-linear saturado conduz a modelos mais parcimoniosos, ditos *Reduzidos* ou *Não Saturados*. Se os dados forem adequadamente descritos por modelos reduzidos, é possível modelar sem perder informação estrutural importante, ganhando-se em simplicidade.

Assim, a análise log-linear tem por objectivo final determinar o modelo interpretável mais parcimonioso que melhor reflicta a estrutura imposta, isto é, o modelo conceptualmente inteligível com menor número de termos, e, portanto, com menos parâmetros, que

possua um bom ajustamento global e parcelar.

4.1 Modelos log-lineares para tabelas bidimensionais

Nesta secção, serão apresentados os modelos log-lineares para tabelas $l \times c$.

4.1.1 Modelo de independência

Duas variáveis categorizadas, organizadas numa tabela bidimensional dizem-se estatisticamente independentes se, para $i = 1, ..., l$ e $j = 1, ..., c$ se tem

$$\pi_{ij} = \pi_{i\bullet} \times \pi_{\bullet j} \tag{4.1}$$

Numa escala logarítmica, a independência em tabelas de contingência bidimensionais assume pois a forma aditiva

$$\log \pi_{ij} = \log \pi_{i\bullet} + \log \pi_{\bullet j} \tag{4.2}$$

Atendendo a que

$$e_{ij} = n\pi_{ij} \Leftrightarrow \log e_{ij} = \log n + \log \pi_{ij} \Leftrightarrow \log \pi_{ij} = \log e_{ij} - \log n \tag{4.3}$$

$$e_{i\bullet} = n\pi_{i\bullet} \Leftrightarrow \log e_{i\bullet} = \log n + \log \pi_{i\bullet} \Leftrightarrow \log \pi_{i\bullet} = \log e_{i\bullet} - \log n \tag{4.4}$$

$$e_{\bullet j} = n\pi_{\bullet j} \Leftrightarrow \log e_{\bullet j} = \log n + \log \pi_{\bullet j} \Leftrightarrow \log \pi_{\bullet j} = \log e_{\bullet j} - \log n \tag{4.5}$$

Então, substituindo em (4.2), obtém-se

$$\log e_{ij} = \log e_{i\bullet} + \log e_{\bullet j} - \log n \tag{4.6}$$

Somando, respectivamente, em i,j e em i e j

$$\left\{ \begin{array}{l} \displaystyle\sum_{i=1}^{l} \log e_{ij} = \sum_{i=1}^{l} \log e_{i\bullet} + l \log e_{\bullet j} - l \log n \\ \displaystyle\sum_{j=1}^{c} \log e_{ij} = c \log e_{i\bullet} + \sum_{j=1}^{c} \log e_{\bullet j} - c \log n \\ \displaystyle\sum_{i=1}^{l} \sum_{j=1}^{c} \log e_{ij} = c \sum_{i=1}^{l} \log e_{i\bullet} + l \sum_{j=1}^{c} \log e_{\bullet j} - cl \log n \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \left\{ \begin{array}{l} \displaystyle\frac{1}{l} \sum_{i=1}^{l} \log e_{ij} = \frac{1}{l} \sum_{i=1}^{l} \log e_{i\bullet} + \log e_{\bullet j} - \log n \\ \displaystyle\frac{1}{c} \sum_{j=1}^{c} \log e_{ij} = \log e_{i\bullet} + \frac{1}{c} \sum_{j=1}^{c} \log e_{\bullet j} - \log n \\ \displaystyle\frac{1}{cl} \sum_{i=1}^{l} \sum_{j=1}^{c} \log e_{ij} = \frac{1}{l} \sum_{i=1}^{l} \log e_{i\bullet} + \frac{1}{c} \sum_{j=1}^{c} \log e_{\bullet j} - \log n \end{array} \right. \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \log e_{\bullet j} = \dfrac{1}{l}\displaystyle\sum_{i=1}^{l} \log e_{ij} - \dfrac{1}{l}\log e_{i\bullet} + \log n \\ \log e_{i\bullet} = \dfrac{1}{c}\displaystyle\sum_{j=1}^{c} c \log e_{ij} - \dfrac{1}{c}\sum_{j=1}^{c} \log e_{\bullet j} + \log n \\ \dfrac{1}{cl}\displaystyle\sum_{i=1}^{l}\sum_{j=1}^{c} \log e_{ij} = \dfrac{1}{l}\sum_{i=1}^{l} \log e_{i\bullet} + \dfrac{1}{c}\sum_{j=1}^{c} \log e_{\bullet j} - \log n \end{cases}$$

Substituindo em (4.6) e adoptando a notação

$$\eta_{ij} = \log e_{ij} \qquad \frac{1}{l}\sum_{i=1}^{l} \eta_{ij} = \bar{\eta}_{\bullet j} \qquad \frac{1}{c}\sum_{j=1}^{c} \eta_{ij} = \bar{\eta}_{i\bullet} \qquad \frac{1}{cl}\sum_{i=1}^{l}\sum_{j=1}^{c} \eta_{ij} = \bar{\eta}_{\bullet\bullet}$$

obtém-se

$$\begin{aligned} \log e_{ij} &= \left(\frac{1}{c}\sum_{j=1}^{c} \log e_{ij} - \frac{1}{c}\sum_{j=1}^{c} \log e_{\bullet j} + \log n\right) \\ &\quad + \left(\frac{1}{l}\sum_{i=1}^{l} \log e_{ij} - \frac{1}{l}\sum_{i=1}^{l} \log e_{i\bullet} + \log n\right) - \log n \\ &= \bar{\eta}_{i\bullet} - \frac{1}{c}\sum_{j=1}^{c} \log e_{\bullet j} + \log n + \bar{\eta}_{\bullet j} - \frac{1}{l}\sum_{i=1}^{l} \log e_{i\bullet} + \log n - \log n \\ &= \bar{\eta}_{i\bullet} + \bar{\eta}_{\bullet j} - \left(\frac{1}{l}\sum_{i=1}^{l} \log e_{i\bullet} + \frac{1}{c}\sum_{j=1}^{c} \log e_{\bullet j} - \log n\right) \\ &= \bar{\eta}_{i\bullet} + \bar{\eta}_{\bullet j} - \bar{\eta}_{\bullet\bullet} \\ &= (\bar{\eta}_{i\bullet} - \bar{\eta}) + (\bar{\eta}_{\bullet j} - \bar{\eta}) + \bar{\eta}_{\bullet\bullet} \end{aligned} \tag{4.7}$$

A equação (4.7) pode, finalmente, escrever-se na forma

$$\log e_{ij} = \mu + \lambda_i^X + \lambda_j^Y \quad \forall\, i, j \tag{4.8}$$

onde

$$\mu = \bar{\eta}_{\bullet\bullet} \tag{4.9}$$

$$\lambda_i^X = \bar{\eta}_{i\bullet} - \bar{\eta}_{\bullet\bullet} \tag{4.10}$$

$$\lambda_j^Y = \bar{\eta}_{\bullet j} - \bar{\eta}_{\bullet\bullet} \tag{4.11}$$

A equação (4.8) representa o *Modelo Log-linear de Independência* para tabelas bidimensionais.

No modelo, tal como no modelo de Análise de Variância a um factor (*One-Way ANOVA*), μ representa o efeito médio global; λ_i^X representa o efeito principal da variável X e λ_j^Y representa o efeito principal da variável Y.

Os parâmetros λ_i^X e λ_j^Y estabelecem que os efeitos principais são medidos pelos desvios relativamente à média global dos logaritmos das frequências esperadas e satisfazem as condições $\sum_{i=1}^{l} \lambda_i^X = 0$ e $\sum_{j=1}^{c} \lambda_j^Y = 0$, existindo $(i-1)$ parâmetros em linha linearmente independentes e $(j-1)$ parâmetros em coluna linearmente independentes.

4.1.2 Modelo saturado

Quando não é possível assumir a independência entre as variáveis, o modelo anterior (equação 4.8) torna-se inadequado, sendo necessário introduzir um termo representativo da interacção entre as variáveis

$$\lambda_{ij}^{XY} = \eta_{ij} - \bar{\eta}_{i\bullet} - \bar{\eta}_{\bullet j} + \bar{\eta}_{\bullet\bullet} \tag{4.12}$$

Considerando a notação adoptada no modelo de independência obtém-se por substituição

$$\log e_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad \forall\, i, j \tag{4.13}$$

Este modelo, designado por *Modelo Saturado*, descreve perfeitamente um conjunto de frequências numa tabela bidimensional. Repare-se que o lado direito da equação é agora análogo ao modelo de Análise de Variância a dois factores (*Two-Way ANOVA*). O parâmetro λ_{ij}^{XY} satisfaz a condição $\sum_{i=1}^{l} \lambda_{ij}^{XY} = \sum_{j=1}^{c} \lambda_{ij}^{XY} = 0$, sendo $(i-1)(j-1)$ destes termos linearmente independentes.

4.1.3 Modelos reduzidos

O modelo de independência e o modelo saturado não são os únicos modelos possíveis de estabelecer em tabelas bidimensionais. Colocando a hipótese de equiprobabilidade das j categorias da variável Y (pelo que X e Y são independentes), tem-se o modelo

$$\log e_{ij} = \mu + \lambda_i^X \quad \forall\, i, j \tag{4.14}$$

Assumindo a equiprobabilidade das i categorias da variável X, tem-se o modelo

$$\log e_{ij} = \mu + \lambda_j^Y \quad \forall\, i, j \tag{4.15}$$

Um último modelo pode obter-se assumido todas as categorias como igualmente prováveis

$$\log e_{ij} = \mu \quad \forall\, i, j \tag{4.16}$$

Estes modelos são designados por não-abrangentes (*noncomprehensive*) por, na sua formulação, não incluírem os efeitos de todas as variáveis. Assim, o estabelecimento de, por exemplo, $\lambda_j^Y = 0$ significa que as frequências associadas a X assumem os mesmos valores para qualquer j. O reconhecimento deste tipo de modelos é importante na medida em que permite reduzir a dimensionalidade da tabela e assim simplificar a análise (o que é particularmente interessante quando se trata de tabelas multidimensionais). Contudo, a análise destes casos não coloca problemas de estimação, pelo que é sobretudo importante atender aos *modelos abrangentes*, isto é, aos modelos que pelo menos contêm os parâmetros relativos aos efeitos principais de cada uma das variáveis. É sobre este modelos que este trabalho incidirá.

Na tabela 4.1 encontram-se resumidos os modelos log-lineares abrangentes para tabelas de contingência bidimensionais.

Tabela 4.1: Modelos log-lineares abrangentes em tabelas de contingência bidimensional

Descrição	*Parâmetros*	*Simbologia*
Saturado	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_{ij}^{XY}$	$[XY]$
Independência	$\mu, \lambda_i^X, \lambda_j^Y$	$[X][Y]$

4.1.4 Interpretação dos parâmetros

Em tabelas bidimensionais, μ simboliza a média dos lc logaritmos dos valores esperados (equação 4.9)). O parâmetro λ_i^X pode interpretar-se como uma medida da variação dos valores estimados relativamente ao valor médio (aumento, quando positivo e diminuição, quando negativo) provocada pela presença da categoria i da variável X (equação 4.10)). Da mesma forma, o parâmetro λ_j^Y dará informação acerca do efeito da categoria j da variável Y (equação 4.11). λ_{ij}^{XY} traduz o desvio de $\log e_{ij}$ relativamente à situação de independência, isto é, fornecem informação relativamente à interacção (de 1ª ordem) entre os níveis de X e Y.

Para melhor entender o significado do parâmetro representativo da interacção entre as variáveis, considere-se o caso mais simples de uma tabela bidimensional, isto é, o caso de uma tabela 2×2. Nesta situação um modelo saturado será aquele que tiver 4 parâmetros. Um modelo possível incluiria os termos $\mu, \lambda_1^X, \lambda_1^Y$ e λ_{11}^{XY}. Nesta situação, tem-se

$$\eta_{11} = \log e_{11} \qquad \eta_{12} = \log e_{12} \qquad \eta_{21} = \log e_{21} \qquad \eta_{22} = \log e_{22}$$

donde

$$\lambda_{11}^{XY} = \eta_{11} - \bar{\eta}_{1\bullet} - \bar{\eta}_{\bullet 1} + \bar{\eta}_{\bullet\bullet} = \frac{\log e_{11} - \log e_{12} - \log e_{21} + \log e_{22}}{4} = \frac{1}{4} \log \left(\frac{e_{11} e_{22}}{e_{12} e_{21}} \right)$$

Equivalentemente, em termos de probabilidades teóricas, tem-se:

$$\lambda_{11}^{XY} = \frac{1}{4}\log\left(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right) = \frac{1}{4}\log\theta$$

.

Assim, o parâmetro λ_{11}^{XY} será nulo quando a razão de produtos cruzados (θ) for igual a 1, isto é, quando as variáveis forem independentes. Quando $\lambda_{11}^{XY} > 0$ há associação positiva e, naturalmente, no caso contrário, há associação negativa entre as variáveis.

Estes resultados são generalizáveis a qualquer conjunto de 4 parâmetros independentes, já que, atendendo às restrições sobre as somas dos parâmetros, se tem (Upton, 1980, página 50)

$$\lambda_2^X = -\lambda_1^X$$

$$\lambda_2^Y = -\lambda_1^Y$$

$$\lambda_{22}^{XY} = -\lambda_{12}^{XY} = -\lambda_{21}^{XY} = \lambda_{11}^{XY}$$

É possível mostrar que os parâmetros de interacção do modelo em tabelas bidimensionais rectangulares são também função da Razão de Produtos Cruzados, tal como exposto para o caso 2×2 (Bishop, Fienberg and Holland, 1980, secção 2.3).

4.2 Modelos log-lineares para tabelas tridimensionais

Frequentemente encontram-se situações onde se pretende analisar as interrelações entre um conjunto de mais de duas variáveis categorizadas, sendo portanto necessário considerar a análise de tabelas multidimensionais. Numa tabela tridimensional, resultante da classificação das variáveis X, Y e Z, podem representar-se as distribuições bivariadas dos vários pares de variáveis para os diferentes níveis da terceira variável, *e.g.* X-Y para os diferentes níveis de Z, através das *Tabelas Parciais* onde Z é controlada, isto é, onde Z é constante. As tabelas de contingência bidimensionais formadas por combinação dos níveis da terceira variável (isto é, por soma das respectivas frequências) são designadas por *Tabelas Marginais* e, ao invés de controlarem Z, ignoram o efeito desta terceira variável. A análise de tabelas parciais pode conduzir a resultados muito distintos da análise de tabelas marginais, podendo mesmo estas análises induzirem a resultados erróneos (Agresti, 1990, página 135). Nesta secção serão abordados os modelos log-lineares adequados à descrição de tabelas tridimensionais.

4.2.1 Modelos de independência

Considere-se, pois, o caso de tabelas tridimensionais, isto é, tabelas onde se regista o número de observações resultantes da classificação cruzada de três variáveis, X, Y e Z,

respectivamente com i, j e k categorias ($i = 1, ..., l; j = 1, ..., c; k = 1, ..., s$). Numa tabela tridimensional, podem estabelecer-se vários tipos de independência entre as variáveis. Representando por π_{ijk} ($i = 1, ..., l; j = 1, ..., c; k = 1, ..., s$) a probabilidade de classificação na cela (i, j, k), isto é

$$\pi_{ijk} = P(X = i, Y = j, Z = k) \tag{4.17}$$

as três variáveis dizem-se *Mutuamente Independentes* quando

$$\pi_{ijk} = \pi_{i\bullet\bullet} \times \pi_{\bullet j\bullet} \times \pi_{\bullet\bullet k} \quad \forall\, i, j, k \tag{4.18}$$

Numa escala logarítmica, tem-se a forma aditiva

$$\log \pi_{ijk} = \log \pi_{i\bullet\bullet} + \log \pi_{\bullet j\bullet} + \log \pi_{\bullet\bullet k} \quad \forall\, i, j, k \tag{4.19}$$

Representando por $\eta_{ijk} = \log e_{ijk}$ ($e_{ijk} > 0,\ \forall\, i, j, k$), então, à semelhança do efectuado para tabelas bidimensionais, pode estabelecer-se

$\mu = \bar{\eta}_{\bullet\bullet\bullet}$

$\lambda_i^X = \bar{\eta}_{i\bullet\bullet} - \bar{\eta}_{\bullet\bullet\bullet}$

$\lambda_j^Y = \bar{\eta}_{\bullet j\bullet} - \eta_{\bullet\bullet\bullet}$

$\lambda_k^Z = \bar{\eta}_{\bullet\bullet k} - \bar{\eta}_{\bullet\bullet\bullet}$

$\lambda_{ij}^{XY} = \bar{\eta}_{ij\bullet} - \bar{\eta}_{i\bullet\bullet} - \bar{\eta}_{\bullet j\bullet} + \bar{\eta}_{\bullet\bullet\bullet}$

$\lambda_{ik}^{XZ} = \bar{\eta}_{i\bullet k} - \bar{\eta}_{i\bullet\bullet} - \bar{\eta}_{\bullet\bullet k} + \bar{\eta}_{\bullet\bullet\bullet}$

$\lambda_{jk}^{YZ} = \bar{\eta}_{\bullet jk} - \bar{\eta}_{\bullet j\bullet} - \bar{\eta}_{\bullet\bullet k} + \bar{\eta}_{\bullet\bullet\bullet}$

$\lambda_{ij}^{XYZ} = \eta_{ijk} - \bar{\eta}_{ij\bullet} - \bar{\eta}_{i\bullet k} - \bar{\eta}_{\bullet jk} + \bar{\eta}_{i\bullet\bullet} + \bar{\eta}_{\bullet j\bullet} + \bar{\eta}_{\bullet\bullet k} - \bar{\eta}_{\bullet\bullet\bullet}$

sendo a soma dos parâmetros relativamente a qualquer índice sempre nula

$$\sum_i \lambda_i^X = \sum_j \lambda_j^Y = \sum_k \lambda_k^Z = \sum_i \lambda_{ij}^{XY} ... = \sum_j \lambda_{ijk}^{XYZ} = \sum_k \lambda_{ijk}^{XYZ} = 0$$

Partindo da equação (4.19) e trabalhando-a algebricamente de modo análogo ao efectuado no caso dos modelos para tabelas bidimensionais, ontem-se o *Model Log-linear de Independência Mútua* em tabelas tridimensionais

$$\log e_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \quad \forall\, i, j, k \tag{4.20}$$

Um segundo tipo de independência entre as variáveis verifica-se quando

$$\pi_{ijk} = \pi_{i\bullet k} \times \pi_{\bullet j\bullet} \quad \forall\, i, j, k \tag{4.21}$$

Esta condição traduz a *Independência Parcial* entre Y e uma nova variável formada pelas $i \times k$ combinações das modalidades das variáveis X e Z sendo, portanto, equivalente à independência em tabelas bidimensionais. O modelo correspondente é designado por *Modelo Log-linear de Independência Parcial entre Y e o par (X,Z)* e é dado por

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} \quad \forall\, i,j,k \tag{4.22}$$

Similarmente, é possível estabelecer os modelos de independência parcial entre X e o par (Y,Z) e entre Z e o par (X,Y), respectivamente

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} \quad \forall\, i,j,k \tag{4.23}$$

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} \quad \forall\, i,j,k \tag{4.24}$$

Repare-se que quando $\pi_{ijk} = \pi_{i\bullet\bullet} \times \pi_{\bullet j\bullet} \times \pi_{\bullet\bullet k} \quad \forall\, i,j,k$ então $\pi_{ijk} = \pi_{i\bullet k} \times \pi_{\bullet j\bullet} = \pi_{ij\bullet} \times \pi_{\bullet\bullet k} = \pi_{\bullet jk} \times \pi_{i\bullet\bullet} \quad \forall\, i,j,k$, ou seja, a independência mútua entre variáveis implica sempre a sua independência parcial.

Por último, considere-se a situação em que duas variáveis são independentes para uma categoria específica da terceira variável, isto é, a terceira variável é controlada. Assim, se X e Y são independentes para a k-ésima categoria de Z, X e Y dizem-se *Condicionalmente Independentes dado o nível k de Z*, isto é

$$\pi_{ij(k)} = \pi_{i\bullet(k)} \times \pi_{\bullet j(k)} \quad \forall\, i,j \tag{4.25}$$

Genericamente, diz-se que existe *Independência Condicional entre X e Y dada Z*, quando a condição anterior se verificar para todos os níveis da variável Z, isto é, para todo o k. O *Modelo Log-linear de Independência Condicional entre X e Y dada Z* é então dado por

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad \forall\, i,j,k \tag{4.26}$$

Da mesma forma, ter-se-ão, respectivamente, os modelos de independência condicional entre X e Z dada Y e entre Y e Z dada X

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} \quad \forall\, i,j,k \tag{4.27}$$

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} \quad \forall\, i,j,k \tag{4.28}$$

Repare-se que a independência parcial é mais forte que a independência condicional, isto é, se existe independência parcial entre, por exemplo, Y e o par (X,Z) então X e Y são condicionalmente independentes (Agresti, 1990, secção 5.2.4).

4.2.2 Modelo sem interacção de 2ª ordem

Quando as três variáveis forem condicionalmente dependentes, então ter-se-à o *Modelo Log-linear sem Interacção de 2ª ordem*

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad \forall\, i,j,k \tag{4.29}$$

4.2.3 Modelo saturado

Usando a nomenclatura adoptada anteriomente, o *Modelo Log-linear Saturado* para tabelas tridimensionais é dado por

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad \forall\, i,j,k \tag{4.30}$$

onde $\lambda_{ij}^{XY}, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ}$ representam as interacções de 1ª ordem e o termo λ_{ijk}^{XYZ} representa a interacção de 2ª ordem.

Os modelos log-lineares abrangentes em tabelas tridimensionais encontram-se resumidos no Tabela 4.2.

Os modelos log-lineares abordados neste texto são *Modelos Hierárquicos*. O principio da hierarquia na construção de modelos log-lineares estabelece que a presença de um determinado termo representativo de um efeito de determinada ordem implica a inclusão no modelo de todos os termos de ordem inferior. É este principio que garante que nestes modelos o conjunto dos termos representativos das interacções entre variáveis representem sempre desvios relativamente à independência[1].

Os modelos não-hierárquicos têm a desvantagem relativamente aos hierárquicos de serem, em regra, analiticamente mais complexos e mais difíceis de interpretar (Paulino e Singer, 2006, páginas 69-70).

[1] Por exemplo, no modelo log-linear saturado para tabelas dimensionais, o principio da hierarquia garante que $\lambda_{ij}^{XY} = \log e_{ij} - \mu - \lambda_i^X \lambda_j^Y$.

4.2.4 Interpretação dos parâmetros

A interpretação dos parâmetros nos modelos log-lineares em tabelas tridimensionais implica a descrição das interacções entre variáveis em termos de razões de produtos cruzados (Agresti, 1990, página 144).

Para interpretar o significado do termo λ_{ijk}^{XYZ}, considere-se, à semelhança do que foi feito para tabelas bidimensionais, o caso mais simples, isto é, a tabela $2 \times 2 \times 2$. Nesta situação, é possível mostrar que (Upton, 1980, página 61; Agresti, 1990, página 145)

$$\begin{aligned}\lambda_{111}^{XYZ} = \lambda_{122}^{XYZ} = \lambda_{221}^{XYZ} = -\lambda_{112}^{XYZ} = -\lambda_{121}^{XYZ} = -\lambda_{211}^{XYZ} = -\lambda_{222}^{XYZ} \\ = \frac{1}{8}\log\left(\frac{\theta_{11(1)}}{\theta_{11(2)}}\right) = \frac{1}{8}\log\left(\frac{\theta_{1(1)1}}{\theta_{1(2)1}}\right) = \frac{1}{8}\log\left(\frac{\theta_{(1)11}}{\theta_{(2)11}}\right)\end{aligned}$$

Estes resultados indicam que a interacção de 2ª ordem será nula quando o quociente entre as razões de produtos cruzados condicionais for 1, isto é, quando a razão de produtos cruzados condicional entre duas variáveis for a mesma para os diferentes níveis da terceira variável. Tal significa que a interacção de segunda ordem mede a diferença de magnitude entre as associações condicionais nas tabelas parciais formadas por uma qualquer partição da tabela tridimensional em tabelas bidimensionais.

Na ausência de interacção de 2ª ordem, os parâmetros que representam as associações entre duas variáveis são proporcionais ao logaritmo da razão de produtos cruzados

Tabela 4.2: Modelos log-lineares abrangentes em tabelas de contingência tridimensionais

Descrição	*Parâmetros*	*Simbologia*
Saturado	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ}, \lambda_{ijk}^{XYZ}$	$[XYZ]$
Sem interacção de 2ª ordem	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ}$	$[XY][XZ][YZ]$
Indep. cond. entre Y e Z, dado X	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{ik}^{XZ}$	$[XY][XZ]$
Indep. cond. entre X e Z, dado Y	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}, \lambda_{jk}^{YZ}$	$[XY][YZ]$
Indep. cond. entre X e Y, dado Z	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ}$	$[XZ][YZ]$
Indep. parcial entre (Y, Z) e X	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{jk}^{YZ}$	$[YZ][X]$
Indep. parcial entre (X, Z) e Y	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ik}^{XZ}$	$[XZ][Y]$
Indep. parcial entre (X, Y) e Z	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z, \lambda_{ij}^{XY}$	$[XY][Z]$
Independência mútua	$\mu, \lambda_i^X, \lambda_j^Y, \lambda_k^Z$	$[X][Y][Z]$

condicional. Isto é, quando $\{\lambda_{ijk}^{XYZ}\} = 0$, então (Agresti, 1990, página 145)

$$\lambda_{11}^{XY} = \frac{1}{4}\log\theta_{11(k)} \quad k = 1, ..., s$$

Assim, no caso das variáveis serem condicionalmente dependentes, os parâmetros representativos da associação dos modelos log-lineares em tabelas tridimensionais podem interpretar-se a partir das tabelas parciais através do calculo da *Razão de Produtos Cruzados Condicional* (equação 2.3).

Existindo independência condicional, por exemplo, entre Z e X ($\theta_{i(j)k} = 1, i = 1, ..., l-1;\ j = 1, ..., c-1;\ k = 1, ..., s$) ou entre Z e Y ($\theta_{(i)jk} = 1, i = 1, ..., l-1;\ j = 1, ..., c-1;\ k = 1, ..., s$), as associações marginal e parcial entre X e Y serão iguais, pelo que o estudo pode ser simplificado através da análise das tabelas marginais[2], recorrendo à *Razão de Produtos Cruzados Marginal* que se define analogamente ao estabelecido para tabelas bidimensionais

$$\theta_{ij} = \frac{\pi_{ij\bullet}\ \pi_{i+1,j+1,\bullet}}{\pi_{i+1,j,\bullet}\ \pi_{i,j+1,\bullet}} \quad i = 1, ..., l-1;\ j = 1, ..., c-1 \tag{4.31}$$

Quando as três variáveis forem condicionalmente dependentes, a análise das tabelas marginais poderá conduzir a conclusões erradas. Este facto está relacionado com o chamado *Paradoxo de Simpson*. Numa tabela 2^3, este paradoxo significa que duas das variáveis podem estar associadas marginalmente num determinado sentido e, simultaneamente, apresentar associação marginal no sentido contrário, quando controladas pela terceira variável (Paulino e Singer, 2006, página 86).

[2]Genericamente, a associação parcial entre duas variáveis será igual à marginal quando a terceira variável for condicionalmente independente de cada uma das duas variáveis associadas. Nesta circunstância a tabela diz-se *desmontável* (*collapsible*) (Agresti, 1990, secção 5.3.3; Paulino e Singer, 2006, secção 4.5). Assim, *e.g.*,

$$\theta_{ij}^{XY} = \theta_{ij(1)} = \theta_{ij(2)} = ... = \theta_{ij(s)} \quad i = 1, ..., l-1;\ j = 1, ..., c-1$$

quando se verificar pelo menos uma das seguintes condições de desmontabilidade (Agresti, 1990, página 146)

$$\theta_{i(j)k} = 1 \quad i = 1, ..., l-1,\ j = 1, ..., c,\ k = 1, ..., s-1$$
$$\theta_{(i)jk} = 1 \quad i = 1, ..., l,\ j = 1, ..., c-1,\ k = 1, ..., s-1$$

Capítulo 5

Estimação paramétrica em modelos log-lineares

Neste capítulo serão apresentados os conceitos relacionados com a estimação de parâmetros em modelos log-lineares. Tal será feito, sem perda de generalidade, tendo por base o modelo probabilístico de Poisson, seguindo Agresti(1990).

5.1 Estatísticas suficientes minímas

Em tabelas bidimensionais as frequências observadas o_{ij} possuem distribuição conjunta de Poisson com função verosimilhança $\mathcal{L}(\{e_{ij}\}|\{o_{ij}\})$ dada pela equação (3.2).

Logaritmizando, obtém-se

$$\log \mathcal{L} = \sum_i \sum_j -e_{ij} + \sum_i \sum_j o_{ij} \log e_{ij} - \sum_i \sum_j \log o_{ij}! \tag{5.1}$$

Dado que o último termo da equação (5.1) não depende de e_{ij} ele pode, para efeitos de estimação paramétrica, ser ignorado obtendo-se o *kernel* da função verosimilhança

$$k(\log \mathcal{L}) = \sum_i \sum_j o_{ij} \log e_{ij} - \sum_i \sum_j e_{ij} \tag{5.2}$$

Atendendo ao modelo (4.13) tem-se que

$$e_{ij} = \exp\{\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}\} \tag{5.3}$$

Substituindo na equação (5.2)

$$\begin{aligned} k(\log \mathcal{L}) &= \sum_i \sum_j o_{ij}(\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) - \sum_i \sum_j \exp(\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) \\ &= n\mu + \sum_i o_{i\bullet}\lambda_i^X + \sum_j o_{\bullet j}\lambda_j^Y + \sum_i \sum_j o_{ij}\lambda_{ij}^{XY} \end{aligned}$$

$$- \sum_i \sum_j \exp(\mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) \tag{5.4}$$

Dado que a Distribuição Poisson pertencente à família exponencial, os coeficientes dos parâmetros λ na equação (5.4) são estatísticas suficientes[1]. Uma vez que que a configuração $\{o_{ij}\}$ implica as anteriores, ela constitui o conjunto de estatísticas suficientes mínimas para o modelo em causa.

Na tabela 5.1 listam-se as estatísticas suficientes suficientes mínimas para os modelos log-lineares abrangentes em tabelas bidimensionais.

Tabela 5.1: Estatísticas suficientes mínimas para os modelos log-lineares abrangentes em tabelas bidimensionais

Modelo	*Estatísticas Suficientes Mínimas*
$[XY]$	$\{o_{ij}\}$
$[X][Y]$	$\{o_{i\bullet}\}, \{o_{\bullet j}\}$

No caso da classificação cruzada de três variáveis, X, Y e Z, procedendo de forma análoga ao descrito para tabelas bidimensionais, obtém-se

$$\begin{aligned} k(\log \mathcal{L}) = n\mu &+ \sum_i o_{i\bullet\bullet}\lambda_i^X + \sum_j o_{\bullet j\bullet}\lambda_j^Y + \sum_k o_{\bullet\bullet k}\lambda_k^Z \\ &+ \sum_i \sum_j o_{ij\bullet}\lambda_{ij}^{XY} + \sum_i \sum_k o_{i\bullet k}\lambda_{ik}^{XZ} + \sum_j \sum_k o_{\bullet jk}\lambda_{jk}^{YZ} \\ &+ \sum_i \sum_j \sum_k o_{ijk}\lambda_{ijk}^{XYZ} \\ &- \sum_i \sum_j \sum_k - \exp(\mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}) \end{aligned} \tag{5.5}$$

Assim, em tabelas tridimensionais, $n, \{o_{i\bullet\bullet}\}, \{o_{\bullet j\bullet}\}, \{o_{\bullet\bullet k}\}, \{o_{ij\bullet}\}, \{o_{i\bullet k}\}, \{o_{\bullet jk}\}$ e $\{o_{ijk}\}$ são estatísticas suficientes no modelo saturado. No caso de modelos re-

[1]Dadas $X_1, ..., X_n$ observações independentes e identicamente distribuídas de uma *fdp* ou *fmp*, $f(x|\tilde{\theta})$, pertencente à família exponencial multiparamétrica, dada por

$$f(x|\tilde{\theta}) = h(x)c(\tilde{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\tilde{\theta})t_i(x)\right)$$

com vector de parâmetros $\tilde{\theta} = (\theta_1, \theta_2, ..., \theta_d),\ d \leq k$, então

$$T(\tilde{X}) = \left(\sum_{j=1}^{n} t_1(X_j), ..., \sum_{j=1}^{n} t_k(X_j)\right)$$

é estatística suficiente para $\tilde{\theta}$ (Casella and Berger, 2002).

duzidos, portanto mais simples que o saturado, alguns termos são nulos (por exemplo, $\lambda_{ijk}^{XYZ} = 0$) e a equação (5.5) simplifica-se. Assim, por exemplo, $n, \{o_{i\bullet\bullet}\}, \{o_{\bullet j\bullet}\}, \{o_{\bullet\bullet k}\}, \{o_{ij\bullet}\}, \{o_{i\bullet k}\}, \{o_{\bullet jk}\}$ são estatísticas suficientes no modelo sem interacção de 2ª ordem. Neste caso, dado que as últimas três configurações $\{o_{ij\bullet}\}, \{o_{i\bullet k}\}, \{o_{\bullet jk}\}$ implicam as anteriores, elas formam o conjunto de estatísticas suficientes mínimas para o modelo em causa.

Na tabela 5.2 listam-se as estatísticas suficientes mínimas para os modelos log-lineares abrangentes em tabelas tridimensionais.

Tabela 5.2: Estatísticas suficientes mínimas para os modelos log-lineares abrangentes em tabelas tridimensionais

Modelo	*Estatísticas Suficientes Mínimas*
$[XYZ]$	$\{o_{ijk}\}$
$[XY][XZ][YZ]$	$\{o_{ij\bullet}\}, \{o_{i\bullet k}\}, \{o_{\bullet jk}\}$
$[XY][XZ]$	$\{o_{ij\bullet}\}, \{o_{i\bullet k}\}$
$[XY][YZ]$	$\{o_{ij\bullet}\}, \{o_{\bullet jk}\}$
$[XZ][YZ]$	$\{o_{i\bullet k}\}, \{o_{\bullet jk}\}$
$[YZ][X]$	$\{o_{\bullet jk}\}, \{o_{i\bullet\bullet}\}$
$[XZ][Y]$	$\{o_{i\bullet k}\}, \{o_{\bullet j\bullet}\}$
$[XY][Z]$	$\{o_{ij\bullet}\}, \{o_{\bullet\bullet k}\}$
$[X][Y][Z]$	$\{o_{i\bullet\bullet}\}, \{o_{\bullet j\bullet}\}, \{o_{\bullet\bullet k}\}$

5.2 Equações de verosimilhança

A determinação dos estimadores de máxima verosimilhança das frequências esperadas implica a maximização da função log verosimilhança.

Assim, por exemplo, para o modelo de independência em tabelas bidimensionais, tem-se

$$\begin{aligned}
\frac{\partial \log \mathcal{L}}{\partial \mu} &= n - \sum_i \sum_j \exp(\mu + \lambda_i^X + \lambda_j^Y) = n - \sum_i \sum_j e_{ij} = n - e_{\bullet\bullet} \\
\frac{\partial \log \mathcal{L}}{\partial \lambda_i^X} &= o_{i\bullet} - \sum_i \sum_j e_{ij} = o_{i\bullet} - e_{i\bullet} \\
\frac{\partial \log \mathcal{L}}{\partial \lambda_j^Y} &= o_{\bullet j} - \sum_i \sum_j e_{ij} = o_{\bullet j} - e_{\bullet j}
\end{aligned}$$

Donde, igualando a zero, obtêm-se as equações de máxima verosimilhança

$$\hat{e}_{\bullet\bullet} = n \tag{5.6}$$

$$\hat{e}_{i\bullet} = o_{i\bullet} \tag{5.7}$$

$$\hat{e}_{\bullet j} = o_{\bullet j} \tag{5.8}$$

No modelo sem interacção de 2ª ordem em tabelas tridimensionais, derivando em ordem aos parâmetros λ representativos das interacções entre variáveis, ter-se-à

$$\hat{e}_{i\bullet\bullet} = o_{i\bullet\bullet} \tag{5.9}$$

$$\hat{e}_{\bullet j\bullet} = o_{\bullet j\bullet} \tag{5.10}$$

$$\hat{e}_{\bullet\bullet k} = o_{\bullet\bullet k} \tag{5.11}$$

As equações (5.9), (5.10) e (5.11), implicam as restantes equações de máxima verosimilhança.

Tal como se constata, as estatísticas suficientes mínimas do modelo coincidem com os estimadores de máxima verosimilhança. Em 1963, Birch demonstrou que este resultado acontece em qualquer modelo log-linear, pelo que uma vez obtidas o conjunto de estatísticas suficientes mínimas de um modelo facilmente se obtêm as estimativas de máxima verosimilhança. Seguidamente enunciam-se os dois resultados de Birch (1963):

Resultado 1 Os totais marginais das frequências observadas pertencentes ao conjunto de estatísticas suficientes mínimas dum modelo são os estimadores de máxima verosimilhança dos totais marginais das frequências esperadas, para um dado λ incluído no modelo.

Resultado 2 Há um único conjunto de estimadores de máxima verosimilhança das frequências esperadas de cada uma das celas da tabela que simultaneamente satisfaz as restrições impostas pelo modelo e as condições do primeiro resultado.

5.3 Métodos de estimação das frequências esperadas

5.3.1 Estimação directa

Nesta sub-secção apresentam-se fórmulas explicitas de determinação das estimativas de máxima verosimilhança das frequências esperadas dos modelos. Quando tal é possível, as estimativas são designadas por *directas* e o método por *Estimação Directa*.

Considere-se o modelo de independência em tabelas bidimensionais. Pegando na equação (5.3) e somando, respectivamente, em i, j e em i e j, obtém-se

$$\sum_i e_{ij} = \sum_i \exp(\mu + \lambda_i^X + \lambda_j^Y) \Leftrightarrow e_{\bullet j} = \exp(\mu + \lambda_j^Y)\left[\sum_i \exp(\lambda_i^X)\right] \tag{5.12}$$

$$\sum_j e_{ij} = \sum_j \exp(\mu + \lambda_i^X + \lambda_j^Y) \Leftrightarrow e_{i\bullet} = \exp(\mu + \lambda_i^X)\left[\sum_j \exp(\lambda_j^Y)\right] \tag{5.13}$$

$$\sum_i \sum_j e_{ij} = \sum_i \sum_j \exp(\mu + \lambda_i^X + \lambda_j^Y) \Leftrightarrow e_{\bullet\bullet} = n = \exp(\mu)\left[\sum_i \exp(\lambda_i^X)\right]\left[\sum_j \exp(\lambda_j^Y)\right] \tag{5.14}$$

Donde se conclui

$$\begin{aligned} \frac{e_{i\bullet} \times e_{\bullet i}}{n} &= \frac{\left(\exp(\mu + \lambda_j^Y)\left[\sum_i \exp(\lambda_i^X)\right]\right) \times \left(\exp(\mu + \lambda_i^X)\left[\sum_j \exp(\lambda_j^Y)\right]\right)}{\exp(\mu)\left[\sum_i \exp(\lambda_i^X)\right]\left[\sum_j \exp(\lambda_j^Y)\right]} \\ &= \exp(\mu + \lambda_i^X + \lambda_j^Y) \\ &= e_{ij} \end{aligned} \tag{5.15}$$

Considerando as equações (5.15), (5.7) e (5.8) e tendo em conta a propriedade da invariância dos EMV, tem-se

$$\hat{e}_{ij} = \frac{\hat{e}_{i\bullet} \times \hat{e}_{\bullet i}}{n} = \frac{o_{i\bullet} \times o_{\bullet i}}{n} \tag{5.16}$$

Pelos resultados de Birch(1963) esta é a única solução de máxima verosimilhança.

A tabela 5.3 resume os EMV das frequências esperadas para os modelos abrangentes em tabelas bidimensionais.

Tabela 5.3: Estimativas de máxima verosimilhança para modelos abrangentes em tabelas bidimensionais

Modelo	$\hat{e}_{ij}$
$[XY]$	o_{ij}
$[X][Y]$	$\frac{o_{i\bullet} \times o_{\bullet i}}{n}$

Procedendo de modo idêntico pode obter-se as EMV em tabelas tridimensionais. A tabela 5.4 resume os resultados em tabelas tridimensionais.

Tabela 5.4: Estimativas de máxima verosimilhança para modelos abrangentes em tabelas tridimensionais

Modelo	$\hat{e}_{ijk}$
$[XYZ]$	o_{ijk}
$[XY][XZ][YZ]$	método iterativo
$[XY][XZ]$	$\frac{o_{ij\bullet} \times o_{i\bullet k}}{n_{i\bullet\bullet}}$
$[XY][YZ]$	$\frac{o_{ij\bullet} \times o_{\bullet jk}}{n_{\bullet j\bullet}}$
$[XZ][YZ]$	$\frac{o_{i\bullet k} \times o_{\bullet jk}}{n_{\bullet\bullet k}}$
$[YZ][X]$	$\frac{o_{\bullet jk} \times o_{i\bullet\bullet}}{n}$
$[XZ][Y]$	$\frac{o_{i\bullet k} \times o_{\bullet j\bullet}}{n}$
$[XY][Z]$	$\frac{o_{ij\bullet} \times o_{\bullet\bullet k}}{n}$
$[X][Y][Z]$	$\frac{o_{i\bullet\bullet} \times o_{\bullet j\bullet} \times o_{\bullet\bullet k}}{n^2}$

5.3.2 Estimação por métodos iterativos

Muitos modelos log-lineares não têm estimativas directas. Assim, por exemplo, no caso do modelo sem interacção de 2ª ordem em tabelas tridimensionais, a estimação das frequências esperadas implica um processo iterativo. Em termos práticos, não é necessário saber quais os modelos que têm estimativas directas e quais os que não têm, porque os métodos iterativos podem ser usados para todos os modelos. Nesta secção resumem-se dois importantes métodos iterativos: *Ajustamento Proporcional Iterativo* e *Método de Newton-Raphson*.

Ajustamento proporcional iterativo

O método de ajustamento proporcional iterativo (Deming and Stephan, 1940) permite obter as estimativas de máxima verosimilhança das frequências esperadas em modelos log-lineares hierárquicos, tendo apenas por base as equações de verosimilhança (Paulino e Singer, 2006, página 291). Este método iterativo começa por usar quaisquer estimativas iniciais, $\{\hat{e}_i^{(0)}\}$, desde que satisfaçam o modelo a ajustar. Multiplicando estes valores por factores de escala apropriados, ajustam-se sucessivamente as estimativas iniciais de modo a que os seus valores coincidam com as frequências marginais que formam as estatísti-

cas suficientes minímas. Este processo prosseguirá até que a variação entre estimativas sucessivas seja desprezável, o que ocorrerá quando todas as equações de verosimilhança estiverem satisfeitas (dentro da aproximação tolerada) (Paulino e Singer, 2006, secção 9.4.2).

Método de Newton-Raphson

Dada uma função $f(\beta)$, o método de Newton-Raphson permite a resolução da equação $f(\beta) = 0$. Pode por isso usar-se para resolver equações como as de verosimilhança que determinam o valor para a qual a função é maximizada. Este método requer uma estimativa inicial, neste caso, para o valor que maximiza a função. O método, inicializado com o valor β_0, define uma sequência de estimativas, $\beta_1, \beta_2, ...$, que converge para um valor $\hat{\beta}$ que satisfaz $f(\hat{\beta}) = 0$. A sequência é definida recursivamente baseando-se o método no desenvolvimento de Taylor da função $f(\hat{\beta})$ em torno de β_0 (Cristensen, 1980, secção 10.5). A estimativa MV do parâmetro será o limite da sucessão de estimativas obtidas pelo método dos mínimos quadrados ponderados, onde os pesos em cada etapa são actualizados em função da estimativa obtida na etapa anterior[2] (Paulino e Singer, 2006, secção 9.4.1).

[2]Este método é, por isso, também designado por Método dos Mínimos Quadrados Iterativamente Reponderados (IRLS - Iteratively Reweighted Least Squares).

Capítulo 6

Avaliação e selecção de modelos

6.1 Ajustamento de modelos

Ao ajustar um modelo log-linear (não saturado), os valores estimados para as frequências são um meio de descrição "suavizada" dos dados, porque os elementos com importância estrutural são retidos e as flutuação devidas ao acaso são descartadas (Bishop, Fienberg and Holland, 1980, página 123). O estudo do ajustamento do modelo aos dados permite avaliar o quão bem o modelo seleccionado descreve os dados. Dado que cada modelo log-linear especifica um determinado tipo de independência, testar o ajustamento de um determinado modelo equivale a testar a hipótese de independência subjacente.

6.1.1 Ajustamento global

Tradicionalmente, há duas medidas que permitem testar o ajustamento global dos modelos log-lineares, isto é, que avaliam a discrepância global entre as estimativas obtidas através do modelo e os valores observados. São elas, a estatística de Pearson

$$X^2 = \sum_i \frac{(o_i - \hat{e}_i)^2}{\hat{e}_i} \quad i = 1, ..., t \tag{6.1}$$

e a estatística razão de verosimilhanças

$$Y^2 = -2 \sum_i o_i \log \frac{\hat{e}_i}{o_i} = 2 \sum_i o_i \log \frac{o_i}{\hat{e}_i} \quad i = 1, ..., t \tag{6.2}$$

sendo t o número total de celas da tabela. Ambas têm distribuição assimptótica χ^2 com um número de g.l. (ν) que varia em função do modelo a ajustar.

Outra estatística de ajustamento, não usada tão frequentemente quanto as anteriores, é a estatística de Freeman-Tukey

$$F = \sum_i \left(\sqrt{o_i} + \sqrt{o_i + 1} - \sqrt{4\hat{e}_i + 1} \right)^2 \quad i = 1, ..., t \tag{6.3}$$

Embora estas sejam, de longe, as estatísticas mais frequentemente usadas (em particular, as estatísticas (6.1) e (6.2), convém salientar que não são as únicas medidas de avaliação de discrepância entre valores observados e valores estimados. Assim, por exemplo, a família de estatísticas definidas por Cressie and Read (1984) e Read and Cressie (1988) (*power divergence statistics*) indexada por um parâmetro real λ

$$\frac{2}{\lambda(\lambda+1)}\sum_i\left[\left(\frac{n_i}{\hat{m}_i}\right)^\lambda-1\right]$$

fornece um conjunto de medidas alternativas para as diferentes concretizações de λ. Estes autores estabeleceram que, para qualquer valor de λ, com n suficientemente grande, a estatística tem distribuição assimptótica χ^2 com ν g.l.. Em particular, para $\lambda=1$ e $\lambda\to 0$ obtêm-se, respectivamente, as estatísticas de Pearson (X^2) e razão de verosimilhanças (Y^2). Considerando $\lambda=-1/2$ e $\lambda\to -1$, a família de estatísticas abrange também a estatística de Freeman-Tukey.

Usando qualquer uma das estatísticas de ajustamento referidas é sempre necessário proceder à determinação do número de graus de liberdade que depende da estrutura dos dados e do número de parâmetros independentes no modelo. Para calcular o número de g.l. associado aos modelos podem usar-se dois métodos (Bishop, Fienberg and Holland, 1980, página 114)

Método 1 O número de g.l. é dado pela diferença entre o número total de celas na tabela de contingência e o número de parâmetros independentes a estimar.
Assim, por exemplo, no modelo $\log e_{ijk}=\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}$ o número estimado de parâmetros independentes é $1+(i-1)+(j-1)+(k-1)+(i-1)(j-1)$, pelo que o número de g.l. associados ao modelo será

$$\nu=ijk-[1+(i-1)+(j-1)+(k-1)+(i-1)(j-1)]=ijk-k+1-ij=(ij-1)(k-1)$$

O número de parâmetros independentes associados aos vários termos dos modelos para tabelas bi- e tridimensionais encontra-se resumido nas tabelas 6.1 e 6.2.

Método 2 O número de g.l. é dado pela soma do número de parâmetros independentes associados aos termos nulos do modelo (isto é, aos termos não incluídos no modelo). No exemplo anterior, ter-se-ia

$$\nu=(i-1)(k-1)+(j-1)(k-1)+(i-1)(j-1)(k-1)=(ij-1)(k-1)$$

Tabela 6.1: Número de parâmetros independentes associados aos termos dos modelos para tabelas bidimensionais

Termo	*Nº de parâmetros independentes*
μ	1
λ_i^X	$i-1$
λ_j^Y	$j-1$
λ_{ij}^{XY}	$(i-1)(j-1)$

Tabela 6.2: Número de parâmetros independentes associados aos termos dos modelos para tabelas tridimensionais

Termo	*Nº de parâmetros independentes*
μ	1
λ_i^X	$i-1$
λ_j^Y	$j-1$
λ_k^Z	$k-1$
λ_{ij}^{XY}	$(i-1)(j-1)$
λ_{ik}^{XZ}	$(i-1)(k-1)$
λ_{jk}^{YZ}	$(j-1)(k-1)$
λ_{ijk}^{XYZ}	$(i-1)(j-1)(k-1)$

Com base nas estatísticas de ajustamento abordadas é possível testar a aderência global dos modelos log-lineares sob as hipóteses

$$H_0 : \pi_i = \pi_{i(0)} \quad vs. \quad H_1 : \pi_i \neq \pi_{i(0)} \quad i = 1, ..., t$$

sendo $\pi_{i(0)} = \hat{e}_i/n$ e $\hat{e}_i$ as estimativas das frequências esperadas sob o modelo. Em particular, as estatísticas X^2 e Y^2 permitem testar a validade das hipóteses, já que, sob H_0, possuem distribuição assimptótica $\chi^2_{(\nu)}$. Deste modo, ao nível de significância α, H_0 será rejeitada quando $X^2_{obs} \geq \chi^2_{\alpha,(\nu)}$ ou, equivalentemente, quando $P(X \geq X^2_{obs}) \leq \alpha,\ X \sim \chi^2_{(\nu)}$.

6.1.2 Ajustamento interno

A análise da qualidade de ajustamento do modelo aos dados não deve restringir-se ao cálculo das estatísticas globais. Em tabelas multidimensionais é particularmente desejável uma análise do ajustamento em cada uma das celas da tabela mediante o cálculo dos *resíduos*. Esta análise é importante na medida em que pode *(1)* explicitar as "causas" de um eventual mau ajustamento global; *(2)* revelar a presença de celas onde o ajuste é pior, e *(3)* permitir identificar celas onde uma maior discrepância entre as frequências observadas e os valores estimados (*outliers*) dite uma influência nefasta no ajustamento global (Paulino e Singer, 2006, página 325).

Adicionalmente, é igualmente importante a análise da magnitude dos *parâmetros* do modelo ajustado, sendo a comparação dos seus valores absolutos bastante útil na interpretação do modelo e na selecção do melhor modelo.

Análise dos resíduos

A quantificação do grau de ajustamento interno é feito através do cálculo dos *resíduos* que se baseiam na análise da distância entre as frequências observadas e as estimativas das frequências esperadas obtidas pelo modelo em causa.

Os *Resíduos de Pearson* (vulgarmente designados por *Resíduos Padronizados*)

$$r_i = \frac{o_i - \hat{e}_i}{\sqrt{\hat{e}_i}} \quad i = 1, ..., t \tag{6.4}$$

são frequentemente usados no estudo do ajustamento interno. Os seus quadrados são as componentes da estatística global de Pearson (equação 6.1). De forma análoga, a razão de verosimilhanças (equação 6.2) pode decompor-se nas componentes $2o_i \log o_i/\hat{e}_i$.

Os resíduos r_i são tais que $\sum_{i=1}^{t} r_i^2$ tem distribuição assimptótica χ^2 com um número de g.l. ν apropriado ao modelo. A estatística $\sum_{i=1}^{t} r_i^2$ possui, portanto, valor médio e variância assimptóticos igual a ν. Como o nº de g.l. (ν) é sempre menor que o nº de

celas (t) da tabela, tal significa que o valor médio assimptótico de cada parcela r_i^2 é, necessariamente, inferior a 1, o que dá uma ideia da ordem de grandeza destes resíduos.

Os resíduos r_i têm distribuição assimptótica normal com valor médio 0 e variância inferior a 1, pelo que a comparação do valor dos resíduos com os quantis da distribuição normal reduzida, permite apenas testar o ajustamento interno de forma conservativa (Agresti, 1990, página 224).

Alternativamente, têm-se os *Resíduos Ajustados*

$$\tilde{r}_i = \frac{o_i - \hat{e}_i}{\sqrt{\hat{e}_i(1-\hat{a}_{ii})}} \quad i = 1, ..., t \tag{6.5}$$

onde $\hat{a}_{ii}$, designada na literatura anglo-saxónica por *leverage*[1], estima a influência da cela i no ajustamento do modelo (Christensen, 1990, página 248). Os resíduos ajustados têm distribuição assimptótica normal com valor médio 0 e variância igual a 1 e permitem, por isso, para grandes amostras, testar se o valor médio dos resíduos é realmente igual a zero. Numa tabela com t celas a realização de t testes aos resíduos implica a utilização do Método de Bonferroni. Assim, um resíduo será significativamente diferente de zero quando

$$|\tilde{r}_i| \geq z_{1-\alpha/2t}$$

sendo z_η o quantil $\eta \times 100\%$ da distribuição normal reduzida e α a probabilidade do erro tipo I global.

Análise dos parâmetros

A análise das estimativas padronizadas dos parâmetros do modelo saturado é importante na medida em que permite: *(1)* ordenar as interacções de acordo com a sua magnitude; *(2)* detectar quando o número de categorias de uma variável pode ser reduzido, e *(3)* ter, *à priori*, uma ideia da qualidade de ajustamento de um modelo não saturado (Bishop, Fienberg and Holland, 1980, página 142). As estimativas padronizadas dos parâmetros calculadas em modelos não saturados fornecem informação importante acerca do efeito das variáveis, respectivas interacções e sua importância relativa.

A comparação dos parâmetros dos termos λ requer o cálculo das estimativas padronizadas $S(\hat{\lambda}) = \hat{\lambda}/\hat{\sigma}(\hat{\lambda})$ sendo $\hat{\sigma}(\hat{\lambda})$ o erro padrão da estimativa aproximado. Good-

[1]Na representação matricial do modelo de regressão $Y = X\beta + \epsilon$, os valores estimados são dados por $\hat{Y} = (X(X^TX)^{-1}X^T)Y = HY$. As *leverages* h_{ii} são os elementos da diagonal da matriz H. Nos modelos log-lineares, as definições são feitas por analogia ao modelo de regressão linear, sendo $\hat{a}_{ii}$ os elementos da diagonal da matriz quadrada $A(\hat{e}) = \{\hat{e}\}A$, onde $\{\hat{e}\}$ é o vector de estimativas das frequências esperadas e $A = X(X^TDX)^{-1}X^TD$ sendo X a matriz de especificação do modelo e $D \equiv D(e) = [d_{ij}]$ com $d_{ii} = e_i,\ d_{ij} = 0, i \neq j$ (Christensen, 1990, secções 6.7 e 10.7).

man (1971) estabelece que os estimadores dos parâmetros devidamente padronizados possuem distribuição assimptótica normal $N(\lambda, 1)$. Sob a hipótese de λ ser 0 então a significância dos parâmetros pode testar-se usando a Distribuição Normal Padrão. Também aqui a realização de múltiplos testes torna aconselhável o uso do Método de Bonferroni.

6.2 Comparação de modelos

Tal como mencionado anteriormente, a análise log-linear tem por objectivo final determinar o modelo mais simples que melhor se ajuste aos dados. Nesta metodologia é por isso fundamental dispor de técnicas que permitam testar modelos mais parcimoniosos contra modelos menos parcimoniosos. À medida que a dimensão da tabela de contingência aumenta, o número de modelos possíveis aumenta também. Assim, é necessário possuir para além das metodologias de comparação, métodos de escolha de modelos.

Considere-se, pois, dois modelos M_a com ν_a g.l. e M_b com ν_b g.l.. Sendo M_a um caso particular do modelo M_b, no sentido em que o último se reduz ao primeiro por remoção de termos, então M_a e M_b dizem-se *encaixados*[2]. A comparação deste tipo de modelos pode ser feita mediante o cálculo das estatísticas

$$X^2(M_a \text{ vs. } M_b) = \sum_i \frac{(\hat{e}_i^{M_b} - \hat{e}_i^{M_a})^2}{\hat{e}_i^{M_a}} \quad i = 1, ..., t \tag{6.6}$$

$$Y^2(M_a \text{ vs. } M_b) = -2 \sum_i \hat{e}_i^{M_b} \log \frac{\hat{e}_i^{M_b}}{\hat{e}_i^{M_a}} \quad i = 1, ..., t \tag{6.7}$$

Embora ambas possuam distribuição assimptótica Qui-quadrado, com $(\nu_b - \nu_a)$ g.l. e, por isso, ambas permitam testar a significância dos termos incluídos em M_b e ausentes em M_a, a estatística Y^2 tem a vantagem de simplificar o processo de comparação dos modelos já que pode ser particionada condicionalmente. Assumindo como termo de comparação o modelo saturado para o qual $\hat{e}_i = o_i$, tem-se

$$Y^2(M_a \text{ vs. } M_{saturado}) = Y^2(M_a) = -2 \sum_i o_i \log \frac{\hat{e}_i^{M_a}}{o_i} \quad i = 1, ..., t \tag{6.8}$$

$$Y^2(M_b \text{ vs. } M_{saturado}) = Y^2(M_b) = -2 \sum_i o_i \log \frac{\hat{e}_i^{M_b}}{o_i} \quad i = 1, ..., t \tag{6.9}$$

pelo que a equação (6.7) equivale a

$$Y^2(M_a \text{ vs. } M_b) = -2 \sum_i o_i \log \frac{\hat{e}_i^{M_b}}{\hat{e}_i^{M_a}} \quad i = 1, ..., t \tag{6.10}$$

[2]Por definição, M_1 diz-se *encaixado* em M_2 se os parâmetros de M_1 são um subconjunto dos parâmetros de M_2 (Paulino e Singer, 2006, página 70).

isto é

$$Y^2(M_a \text{ vs. } M_b) = Y^2(M_a \text{ vs. } M_{saturado}) - Y^2(M_b \text{ vs. } M_{saturado}) \tag{6.11}$$

ou, simplificando a notação

$$Y^2(M_a|M_b) = Y^2(M_a) - Y^2(M_b) \tag{6.12}$$

A estatística $Y^2(M_a|M_b)$ será tanto mais elevada quanto melhor for o ajustamento de M_a relativamente ao conseguido com M_b. Ela possui distribuição assimptótica qui-quadrado com $(\nu_b - \nu_a)$ g.l., pelo que o termo incluído em M_b e ausente em M_a deverá ser incluído no modelo quando

$$Y^2(M_a|M_b) \geq \chi_{\alpha;(\nu_b-\nu_a)}$$

sendo χ_η o quantil $\eta \times 100\%$ da distribuição qui-quadrado e α a probabilidade do erro tipo I ou nível de significância do teste.

Considere-se agora, em detalhe, alguns testes relevantes na comparação de modelos log-lineares. Em tabelas bidimensionais, a estatística (6.6) permite, através da comparação dos modelos encaixados $[XY]$ *vs.* $[X][Y]$, testar as hipóteses $H_0 : \lambda_{ij}^{XY} = 0$ *vs.* $H_1 : \lambda_{ij}^{XY} \neq 0$. Repare-se que este procedimento equivale, de facto, a testar a hipótese de independência entre variáveis expressa por $H_0 : \pi_{ij} = \pi_{i\bullet} \times \pi_{\bullet j}$.

Em tabelas tridimensionais, a comparação dos modelos encaixados $[XYZ]$ *vs.* $[XY][XZ][YZ]$ permite testar a significância da interacção de 2ª ordem expressa nas hipóteses $H_0 : \lambda_{ijk}^{XYZ} = 0$ *vs.* $H_1 : \lambda_{ijk}^{XYZ} \neq 0$. A rejeição de H_0 leva a concluir que o modelo saturado é aquele que descreve adequadamente os dados. Ao contrário, a não rejeição de H_0 conduz, na procura do modelo significativo mais parcimonioso, a testar as hipóteses de independência condicional

$$H_0 : \pi_{(i)jk} = \pi_{(i)\bullet k} \times \pi_{(i)j\bullet} \Leftrightarrow H_0 : \lambda_{jk}^{YZ} = 0$$

$$H_0 : \pi_{i(j)k} = \pi_{i(j)\bullet} \times \pi_{\bullet(j)\bullet} \Leftrightarrow H_0 : \lambda_{ik}^{XZ} = 0$$

$$H_0 : \pi_{ij(k)} = \pi_{i\bullet(k)} \times \pi_{\bullet j(k)} \Leftrightarrow H_0 : \lambda_{ij}^{XY} = 0$$

comparando, respectivamente, os modelos encaixados $[XY][XZ][YZ]$ *vs.* $[XY][XZ]$, $[XY][XZ][YZ]$ *vs.* $[XY][YZ]$ e $[XY][XZ][YZ]$ *vs.* $[XZ][YZ]$.

A não rejeição de todas as hipóteses de independência condicional implica a independência completa entre as variáveis. Contrariamente, a rejeição de todas as hipóteses de independência condicional permite concluir que o modelo sem interacção de 2ª ordem é o adequado. Nas situações intermédias é ainda necessário testar a independência parcial entre variáveis. Neste caso as hipóteses possíveis são

$$H_0 : \pi_{ijk} = \pi_{\bullet jk} \times \pi_{i\bullet\bullet} \Leftrightarrow H_0 : \lambda_{ij}^{XY} = \lambda_{ik}^{XZ} = 0$$

$$H_0 : \pi_{ijk} = \pi_{i\bullet k} \times \pi_{\bullet j\bullet} \Leftrightarrow H_0 : \lambda_{ij}^{XY} = \lambda_{jk}^{YZ} = 0$$

$$H_0 : \pi_{ijk} = \pi_{ij\bullet} \times \pi_{\bullet\bullet k} \Leftrightarrow H_0 : \lambda_{ik}^{XZ} = \lambda_{jk}^{YZ} = 0$$

A realização de testes sob as hipóteses anteriores permitirá decidir sobre a adequação dos modelos de independência parcial *vs* o modelo de independência completa.

Note-se que na utilização da partição condicional de Y^2 para comparação de mais de dois modelos devem usar-se níveis de significância corrigidos por forma a garantir que o erro de 1ª espécie não excede o valor α.

Os procedimentos de comparação de modelos até aqui descritos são exclusivamente aplicáveis a modelos encaixados. Contudo, a estatística Y^2 pode também ser usada para comparar modelos não encaixados, usando o *Critério de Informação de Akaike*

$$AIC(M) = -2(Y^2(M) - p) \tag{6.13}$$

sendo p o número de parâmetros no modelo. O primeiro termo é uma medida do quão bem o modelo M se ajusta aos dados e o segundo termo corresponde à penalização pela complexidade do modelo (número de parâmetros). Subtraindo a esta quantidade o AIC relativo ao modelo saturado, obtém-se a *Medida de Informação de Akaike*

$$A(M) = Y^2(M) - 2\nu \tag{6.14}$$

Um modelo será tanto melhor quanto menor forem os valores de AIC e A.

Outra medida alternativa é o *Critério de Informação de Bayes*

$$BIC(M) = -2Y^2(M) + p \log n \tag{6.15}$$

equivalente, para efeito de comparação de modelos, à medida

$$B(M) = Y^2(M) - \log n\nu \tag{6.16}$$

Estas medidas têm directamente em conta a dimensão amostral. Comparando os critérios AIC e BIC, verifica-se que o decréscimo do critério de Bayes é mais lento que o verificado no critério de Akaike para modelos mais complexos, à medida que n aumenta (Agresti, 1990, página 251).

6.3 Selecção de modelos

Dado o elevado número de modelos log-lineares em tabelas multidimensionais, em particular, para um número de variáveis superior ou igual a 4, é necessário dispor de métodos

de selecção do melhor modelo, já que é impraticável ajustar todos é avaliar *à posteriori* qual o melhor[3].

Neste contexto, entende-se como melhor modelo aquele que, apresentando um bom ajustamento aos dados, seja interpretável e possua um número mínimo de parâmetros (o mais parcimonioso).

6.3.1 Selecção do modelo inicial

Alguns estudos são desenhados para responder a questões específicas. Quando tal acontece, as questões teóricas são o guia para escolher os potenciais termos dos modelos e as análises confirmatórias responderão assim às questões colocadas testando um número restrito de modelos.

Contudo, em muitas situações os estudos são eminentemente exploratórios e implicam, por isso, a análise de um conjunto alargado de modelos. Neste caso, para escolher o melhor modelo existem procedimentos algorítmicos (*Stepwise Procedures*) que, mediante um conjunto de regras, ditam a adição ou a remoção de termos de um modelo inicial com vista à obtenção do modelo final.

Coloca-se, portanto, antes de mais, a questão da selecção do modelo inicial. Para tal existem várias estratégias possíveis (Christensen, 1990, secção 6.2; Agresti, 1990, secção 7.2.3; Paulino e Singer, 2006, secção 9.6.1).

Uma primeira possibilidade consiste em escolher o modelo mais complexo dentro de uma determinada ordem. Este procedimento consiste na escolha do modelo inicial comparando os modelos de independência, de interacção de 1ª ordem, de interacção de 2ª ordem, e assim sucessivamente. Os teste de ajustamento descritos anteriormente permitem então identificar o modelo mais complexo que não se ajusta aos dados. Neste caso este será o modelo inicial podendo, em seguida, considerar-se a remoção de termos. Alternativamente, o modelo inicial pode ser o modelo mais simples que melhor se ajusta aos dados procedendo-se em seguida à adição de termos.

Uma segunda possibilidade consiste na análise da significância de cada um dos termos incluídos no modelo saturado, através dos testes de associação parcial e marginal sugeridos por Brown (1976) e Benedetti e Brown (1978). De acordo com estes autores, a significância da associação parcial relativa a um termo pode analisar-se testando o anulamento condicional desse termo no quadro de um modelo hierárquico contendo todos os termos da mesma ordem daquele. Assim, a independência condicional entre X e Y, dada Z pode testar-se mediante o cálculo da estatística $Y^2([XZ][YZ] \mid [XY][XZ][YZ])$ sob a

[3]Estes procedimentos são particularmente importantes no caso da classificação cruzada de 4 ou mais variáveis. Na prática, em tabelas tridimensionais, dado o número ainda relativamente restrito de modelos, analisam-se todos os modelos

hipótese $H_0 : \lambda_{ij}^{XY} = 0,\ \forall\, i, j$.

A associação marginal de um termo é analisada testando o anulamento condicional desse termo no quadro do modelo hierárquico que contém esse termo como o parâmetro mais complexo. Assim, a independência parcial entre (X, Y) e Z pode testar-se mediante o cálculo da estatística $Y^2([X][Y][Z] \mid [XY][Z])$ sob a hipótese $H_0 : \lambda_{ij}^{XY} = 0,\ \forall\, i, j$.

Para identificar os termos não significativamente diferentes de zero pode ainda usar-se outro método (Goodman, 1971) baseado no facto da estatística $(\hat{\lambda} - \lambda_0)/\hat{\sigma}(\hat{\lambda})$ possuir, sob $H_0 : \lambda = \lambda_0$, distribuição assimptótica normal[4].

6.3.2 Métodos *stepwise*

Uma vez escolhido o modelo inicial por uma das metodologias referidas na secção anterior pode iniciar-se um processo de selecção estatística (*stepwise*) do melhor modelo, adicionando ou removendo sequencialmente termos ao modelo inicial.

Existem três tipos principais de procedimentos *stepwise*:

Selecção Progressiva: Neste método são adicionados termos ao modelo inicial sempre que a sua introdução melhore significativamente o ajustamento aos dados. Em cada passo, o termo a incluir será aquele que conduza ao modelo com menor valor-p em termos da estatística Y^2 condicional, tendo como limite um nível de significância máximo (α_E), até ao qual se aceita a entrada do termo no modelo. Em geral, o valor α_E assume os valores 0.05 ou 0.10. Repare-se que a utilização de níveis de significância mais elevados que os usuais (isto é, superiores a 0.10), por um lado, aumenta a probabilidade de rejeitar a hipótese nula quando ela é verdadeira (o que corresponde a aumentar a probabilidade de decidir que um termo é significativo quando na realidade não é), mas, por outro lado, diminui a probabilidade não rejeitar a hipótese nula quando ela é falsa, isto é, diminui a probabilidade não encontrar um termo que é significativo. Assim, parece preferível utilizar um nível de significância para entrar mais elevado, *e.g.*, entre 0.10 e 0.20, já que valores mais baixos podem excluir termos que sejam relevantes na estimação das contagens. O processo de inclusão de termos cessa quando o modelo deixa de ser estatisticamente significativo ou quando se esgotem os termos excluídos;

Selecção Regressiva: Neste método são removidos termos ao modelo inicial sempre que a sua remoção melhore significativamente o ajustamento aos dados. A exclusão dos termos é feita com base nos resultados da comparação de modelos. Em cada passo, o termo a ser excluído é aquele para o qual se obteve a maior probabilidade

[4]Estes tipo de testes baseados numa estatística $Z_n = (W_n - \theta_0)/S_n$, sendo W_n um estimador de θ e S_n o desvio-padrão aproximado do estimador W_n e sob $H_0 : \theta = \theta_0$, são designados por Testes de Wald.

de significância (valor-p) dos testes de ajustamento condicional. Neste método, a eliminação/permanência de um determinado termo no modelo depende do nível de significância mínimo (α_R), até ao qual se aceita a permanência do termo no modelo. O processo termina quanto todos os termos presentes no modelo contribuírem significativamente para a estimação das contagens, isto é, quando $p < \alpha_R$ (sendo α_R usualmente fixado em 0.05 ou 0.10).

Selecção Mista: Este último procedimento é um compromisso entre os dois anteriores, ocorrendo uma selecção progressiva com eliminação regressiva. Os valores de α para adicionar e remover termos normalmente utilizados são os mencionados anteriormente, com $\alpha_E < \alpha_R$.

Capítulo 7

Modelos log-lineares ordinais

Os modelos log-lineares abordados até aqui tratam todas as variáveis como nominais, no sentido em que as estimativas dos parâmetros e as estatísticas de ajustamento são invariantes em relação à ordem das categorias. Consequentemente, quando pelo menos uma das variáveis é ordinal, estes modelos ignoram informação importante contida nos dados. Nestas situações, a modelação da associação entre variáveis ordinais através dos modelos log-lineares deve ser adaptada de modo a reflectir características como a monotonia e a ordem estocástica dos dados.

A adaptação dos modelos log-lineares a variáveis ordinais implica a atribuição de *scores* às categorias das variáveis. A definição dos *scores* afecta a interpretação dos termos do modelo que reflectem a ordinalidade das variáveis. Na prática, a atribuição dos *scores* é feita de modo a simplificar a interpretação dos modelos. Neste sentido é frequente usar *scores* igualmente espaçados[1]. Na maioria das situações, são escolhidas distâncias unitárias sendo os *scores* definidos pelos próprios índices das categorias ordenadas crescentemente[2]. Outra possibilidade, menos frequente, consiste em trabalhar com *scores* de soma nula[3]. Quando a atribuição dos *scores* não surge de forma natural, deve efectuar-se uma análise de sensibilidade, ou seja, é aconselhável a atribuição de *scores* de vários modos, para que se possa verificar se as conclusões dependem substancialmente dos *scores* atribuídos (Agresti, 1990, secção 8.6.2).

Neste trabalho abordar-se-à apenas a situação em que todas as variáveis envolvidas na classificação cruzada dos dados são ordinais embora, naturalmente, possam ser consideradas as situações intermédias em que apenas algumas das variáveis em estudo são ordinais sendo as restantes nominais.

[1]Em tabelas bidimensionais $u_{i+1} - u_i = a$ e $v_{j+1} - v_j = b$ e de forma análoga em tabelas multidimensionais.

[2]Em tabelas bidimensionais $u_i = i$ e $v_j = j$ e de forma análoga em tabelas multidimensionais.

[3]Em tabelas bidimensionais $\{u_i = i - (l+1)/2\}$ e $\{v_j = j - (c+1)/2\}$ e de forma análoga em tabelas multidimensionais.

7.1 Modelos log-lineares ordinais em tabelas bidimensionais

Considere-se, pois, em tabelas bidimensionais, os *scores* $\{u_i\}$ e $\{v_j\}$ $(i = 1, ..., l; j = 1, ..., c)$, relativos às categorias representadas em linha e coluna, respectivamente, tal que $u_1 \leq u_2 \leq ... \leq u_l$ e $v_1 \leq v_2 \leq ... \leq v_c$. Neste caso, o modelo log-linear ordinal pode escrever-se na forma

$$\log e_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta^{XY}(u_i - \bar{u})(v_j - \bar{v}) \quad \forall\, i, j \tag{7.1}$$

onde $\bar{u} = \frac{1}{l}\sum_i u_i$, $\bar{v} = \frac{1}{c}\sum_j v_j$ e $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$.

O termo $\beta^{XY}(u_i - \bar{u})(v_j - \bar{v})$ reflecte a natureza ordinal de X e Y, através dos respectivos *scores*. Como $\{u_i\}$ e $\{v_j\}$ são fixos, este modelo só tem mais um parâmetro — β — que o modelo geral de independência, pelo que o número de g.l. é $ij - [1 + (i-1) + (j-1) + 1] = ij - i - j$.

Repare-se que o modelo de independência geral em tabelas bidimensionais é um caso particular do modelo anterior, quando $\beta = 0$. Tal significa que $\beta^{XY}(u_i - \bar{u})(v_j - \bar{v})$ representa o desvio de $\log e_{ij}$ relativamente à situação de independência, ou seja, os desvios à independência são função linear dos respectivos scores centrados. Este modelo, pode, por isso, ser designado por *Modelo de Associação Linear por Linear*.

Por outro lado, o modelo (7.1) não é mais que um caso especial do modelo saturado em tabelas bidimensionais (equação 4.13) onde o termo λ_{ij}^{XY} assume a forma estruturada $\beta^{XY}(u_i - \bar{u})(v_j - \bar{v})$.

O parâmetro β descreve a magnitude da associação entre as variáveis X e Y. Se $\beta > 0$, $\beta^{XY}(u_i - \bar{u})(v_j - \bar{v})$ é positivo e as frequências esperadas para pequenos e grandes valores de X e Y (isto é, nas celas dos cantos superior esquerdo e inferior direito da tabela de contingência) são, consequentemente, superiores às obtidas no caso de independência. Se $\beta < 0$, as frequências esperadas para grandes valores de X e pequenos valores de Y e vice-versa, serão menores que no caso de independência. Este modelo é, portanto, um modo de impor uma tendência monotónica com direcção dependente do sinal de β.

O modelo (7.1) quando aplicado com *scores* igualmente espaçados é designado por *Modelo de Associação Uniforme*. Esta designação deriva do facto de, sob este modelo, as Razões de Chance Locais serem todas iguais. De facto, escolhidos os *scores* tal que $u_{i+1} - u_i = a$ $(i = 1, ..., l-1)$ e $v_{j+1} - v_j = b$ $(j = 1, ..., c-1)$ então

$$\begin{aligned}
\log \theta_{ij} &= \log\left(\frac{e_{i,j}\ e_{i+1,j+1}}{e_{i+1,j}\ e_{i,j+1}}\right) \\
&= \log e_{i,j} + \log e_{i+1,j+1} - \log e_{i+1,j} - \log e_{i,j+1} \\
&= \beta(u_i - \bar{u})(v_j - \bar{v}) + \beta(u_{i+1} - \bar{u})(v_{j+1} - \bar{v})
\end{aligned}$$

$$
\begin{aligned}
& -\beta(u_i-\bar{u})(v_{j+1}-\bar{v})-\beta(u_{i+1}-\bar{u})(v_j-\bar{v}) \\
= \;& \beta\Big\{(u_i-\bar{u})\Big[(v_j-\bar{v})-(v_{j+1}-\bar{v})\Big]+(u_{i+1}-\bar{u})\Big[(v_{j+1}-\bar{v})-(v_j-\bar{v})\Big]\Big\} \\
= \;& \beta\Big[(u_{i+1}-u_i)(v_{j+1}-v_j)\Big] \\
= \;& \beta ab \qquad (7.2)
\end{aligned}
$$

donde para *scores* definidos pelos próprios índices ordinais $\{u_i = i\}$ e $\{v_j = j\}$ e espaçamentos unitários ($a = b = 1$), todas as Razões de Chances, sob o modelo, são simplesmente iguais a $\exp(\beta)$.

7.1.1 Estimação do modelo

Considere-se a equação (5.2) que estabelece a função kernel log-verosimilhança. Sob o modelo (7.1), obtém-se por substituição

$$
\begin{aligned}
k(\log\mathcal{L}) \;=\; & \sum_{i,j} o_{ij}\Big[\mu+\lambda_i^X+\lambda_j^Y+\beta(u_i-\bar{u})(v_j-\bar{v})\Big] \\
& -\sum_{i,j}\exp\Big[\mu+\lambda_i^X+\lambda_j^Y+\beta(u_i-\bar{u})(v_j-\bar{v})\Big] \\
=\; & n\mu+\sum_i o_{i\bullet}\lambda_i^X+\sum_j o_{\bullet j}\lambda_j^Y+\beta\sum_i\sum_j o_{ij}(u_i-\bar{u})(v_j-\bar{v}) \\
& -\sum_i\sum_j\exp\Big[\mu+\lambda_i^X+\lambda_j^Y+\beta(u_i-\bar{u})(v_j-\bar{v})\Big] \qquad (7.3)
\end{aligned}
$$

Igualando a zero as derivadas parciais em ordem a λ_i^X, λ_j^Y e β (tal como efectuado na secção 5.2) obtêm-se, respectivamente, as equações de máxima verosimilhança

$$\hat{e}_{i\bullet} = o_{i\bullet}$$

$$\hat{e}_{\bullet j} = o_{\bullet j}$$

$$\sum_{i=1}^{l}\sum_{j=1}^{c}(u_i-\bar{u})(v_j-\bar{v})\hat{e}_{ij} = \sum_{i=1}^{l}\sum_{j=1}^{c}(u_i-\bar{u})(v_j-\bar{v})o_{ij}$$

sendo a última equivalente a

$$\sum_{i=1}^{l}\sum_{j=1}^{c}(u_i-\bar{u})(v_j-\bar{v})\hat{\pi}_{ij} = \sum_{i=1}^{l}\sum_{j=1}^{c}(u_i-\bar{u})(v_j-\bar{v})p_{ij}$$

com $\hat{\pi}_{ij} = \hat{e}_{ij}/n$ e $p_{ij} = o_{ij}/n$.

Estas equações não têm soluções directas, podendo obter-se as EMV usando um método iterativo, *e.g.*, o Método de Newton-Raphson. Em termos analíticos, o parâmetro de maior importância é β. Para grandes amostras, $\hat{\beta}$ possui distribuição assimptótica normal, com erro padrão estimado pelo inverso da estimativa da matriz de informação (Agresti, 1990, página 267).

7.2 Modelos log-lineares ordinais em tabelas tridimensionais

Em tabelas multidimensionais, os métodos para considerar a ordinalidade das variáveis são basicamente os mesmos expostos para tabelas bidimensionais.

Em particular, em tabelas tridimensionais, os modelos ordinais sem interacção de 2ª ordem são casos especiais do modelo $[XY][XZ][YZ]$. Um modelo parcimonioso que comtemple a ordem das categorias é, também neste caso, obtido mediante a substituição dos termos λ por termos estruturados que incluem *scores* ordinais. Assim, na modelação das frequências de uma tabela $l \times c \times s$, onde as categorias ordinais são representadas pelos scores $\{u_i\}$, $\{v_j\}$ e $\{w_k\}$, os termos λ_{ij}^{XY}, λ_{ik}^{XZ} e λ_{jk}^{YZ} podem ser substituídos, obtendo-se o modelo

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta^{XY}(u_i - \bar{u})(v_j - \bar{v}) + \beta^{XZ}(u_i - \bar{u})(w_k - \bar{w}) + \beta^{YZ}(v_j - \bar{v})(w_k - \bar{w}) \quad \forall\, i, j, k \tag{7.4}$$

Os parâmetros β^{XY}, β^{XZ} e β^{YZ} descrevem as associações parciais entre as variáveis. Quando $\beta = 0$ as correspondentes variáveis serão condicionalmente independentes. Como este modelo tem mais 3 parâmetros que o modelo de independência em tabelas tridimensionais, possui $ijk - i - j - k - 1$ g.l..

No caso de espaçamentos unitários entre *scores*, então, para qualquer nível k fixo de Z, a Razão de Chances Local é igual a $\exp(\beta^{XY})$; para qualquer nível j fixo de Y, a Razão de Chances Local é igual a $\exp(\beta^{XZ})$ e para qualquer nível i fixo de X, a Razão de Chances Local é igual a $\exp(\beta^{YZ})$, isto é

$$\begin{aligned} \log\left(\frac{e_{i,j,k}\; e_{i+1,j+1,k}}{e_{i,j+1,k}\; e_{i+1,j,k}}\right) &= \beta^{XY} \\ \log\left(\frac{e_{i,j,k}\; e_{i+1,j,k+1}}{e_{i,j,k+1}\; e_{i+1,j,k}}\right) &= \beta^{XZ} \\ \log\left(\frac{e_{i,j,k}\; e_{i,j+1,k+1}}{e_{i,j+1,k}\; e_{i,j,k+1}}\right) &= \beta^{YZ} \end{aligned} \tag{7.5}$$

De acordo com (7.5) a Razão de Chances Locais é uniforme para cada par de variáveis e a magnitude da associação entre duas variáveis é homogénea ao longo dos níveis da terceira variável. É por esta razão que estes modelos são designados por *Modelos de Associação Uniforme Homogénea* (Christensen, 1990, página 267).

No caso dum modelo sem interacção de 2ª ordem não ser adequado para uma boa descrição dos dados, podem ajustar-se modelos com interacção de 2ª ordem que, ao contrário do modelo $[XYZ]$, não são saturados. Os modelos

$$\log e_{ijk} = \mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XY}+\lambda_{ik}^{XZ}+\lambda_{jk}^{YZ}+\beta^{XYZ}(u_i-\bar{u})(v_j-\bar{v})(w_k-\bar{w}) \quad \forall\, i,j,k \tag{7.6}$$

e

$$\begin{aligned}\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta^{XY}(u_i-\bar{u})(v_j-\bar{v}) + \beta^{XZ}(u_i-\bar{u})(w_k-\bar{w}) \\ +\beta^{YZ}(v_j-\bar{v})(w_k-\bar{w}) + \beta^{XYZ}(u_i-\bar{u})(v_j-\bar{v})(w_k-\bar{w}) \quad \forall\, i,j,k\end{aligned} \tag{7.7}$$

incluem a interacção de 2ª ordem e lidam com as interacções de 1ª ordem de diferentes modos. Para estes modelos, no caso de espaçamentos unitários entre *scores*, procedendo analogamente ao efectuado em (6.2), mostra-se que (Agresti, 1984, página 95)

$$\log \theta_{ijk} = \beta^{XYZ} \tag{7.8}$$

$$\log \theta_{ij(k)} = \beta^{XY} + \beta^{XYZ}\left[k - \frac{s+1}{2}\right] \tag{7.9}$$

A equação (7.8) revela que a interacção local é constante para todas as sub-tabelas $2 \times 2 \times 2$ formadas por linhas, colunas e estratos adjacentes. Por seu turno, a equação (7.9) mostra que a Razão de Chances Locais é uniforme para cada par de variáveis dentro de um determinado nível da terceira variável, mas a magnitude da associação entre duas variáveis varia linearmente ao longo dos níveis da terceira variável. Os modelos log-lineares ordinais com interacção de 2ª ordem são, por isso, ser descritos como *Modelos de Associação Uniforme Heterogénea* (Christensen, 1990, página 267).

7.2.1 Estimação do modelo

A estimação do modelo, no caso de tabelas tridimensionais, é feita de modo análogo ao exposto para tabelas bidimensionais.

No caso do modelo sem interacção de 2ª ordem que considera a ordinalidade das três variáveis (equação (7.4)), é possível mostrar que, procedendo de forma análoga ao efectuado na secção anterior, se têm as equações de verosimilhança

$$\sum_{i=1}^{l}\sum_{j=1}^{c}(u_i-\bar{u})(v_j-\bar{v})\hat{e}_{ij\bullet} = \sum_{i=1}^{l}\sum_{j=1}^{c}(u_i-\bar{u})(v_j-\bar{v})o_{ij\bullet} \tag{7.10}$$

$$\sum_{i=1}^{l}\sum_{k=1}^{s}(u_i-\bar{u})(w_k-\bar{w})\hat{e}_{i\bullet k} = \sum_{i=1}^{l}\sum_{k=1}^{s}(u_i-\bar{u})(w_k-\bar{w})o_{i\bullet k} \tag{7.11}$$

$$\sum_{j=1}^{c}\sum_{k=1}^{s}(v_j-\bar{v})(w_k-\bar{w})\hat{e}_{\bullet jk}=\sum_{j=1}^{c}\sum_{k=1}^{s}(v_j-\bar{v})(w_k-\bar{w})o_{\bullet jk} \tag{7.12}$$

Sob o modelo definido em (7.6), as equações de verosimilhança serão as do modelo (não ordinal) sem interacção de 2ª ordem (equação (4.29)) acrescidas da equação

$$\sum_{i=1}^{l}\sum_{j=1}^{c}\sum_{k=1}^{s}(u_i-\bar{u})(v_j-\bar{v})(w_k-\bar{w})\hat{e}_{ijk}=\sum_{i=1}^{l}\sum_{j=1}^{c}\sum_{k=1}^{s}(u_i-\bar{u})(v_j-\bar{v})(w_k-\bar{w})o_{ijk} \tag{7.13}$$

Para o modelo (7.7) as equações de verosimilhança serão as formadas pelo conjunto (7.10) a (7.13).

As EMV dos parâmetros podem obter-se, também neste caso, usando, por exemplo, o Método de Newton-Raphson.

7.3 Selecção de modelos log-lineares ordinais

A introdução de termos de interacção que incluem os *scores* ordinais, aumenta muito o número de possíveis modelos a ajustar. É, por isso, de grande importância dispor de estratégias de escolha do melhor modelo.

Agresti (1990, página 293) sugere a estratégia de ter como guia os modelos log-lineares sem considerar a ordinalidade das variáveis. Uma vez obtido o modelo log-linear que se ajuste bem aos dados, procede-se à substituição faseada dos termos de interacção por termos estruturados baseados nos *scores* .

Capítulo 8

Zeros amostrais e estruturais

Na análise de tabelas de contingência é frequente existiram celas com frequências observadas nulas. Este tipo de celas podem classificar-se em dois grupos: *Zeros estruturais* ou *Zeros fixos* e *Zeros amostrais* ou *Zeros aleatórios*.

Os zeros estruturais correspondem a celas nas quais é impossível registar observações. Por seu turno, os zeros amostrais correspondem a celas com frequências nulas mas onde, no entanto, é possível obter contagens não nulas. As tabelas com zeros estruturais são designadas por *Tabelas Incompletas*.

Na análise de tabelas de contingência com zeros há situações onde as celas podem, e devem, ser divididas em subconjuntos separando-se celas nulas de celas não nulas. Nestes casos, a tabela original diz-se *separável* (Bishop, Fienberg and Holland, 1980, secções 5.2.2 e 5.4.2; Upton, 1980, secção 9.3). A análise de tabelas incompletas separáveis é mais simples do que a análise de tabelas inseparáveis, já que o processo de separação conduz a sub-tabelas completas que podem, para efeitos de estimação e ajustamento de modelos, ser analisadas individualmente[1]. Assim, e apesar de em muitas situações as tabelas contendo zeros serem tabelas inseparáveis, deve, num primeiro passo, dada a maior facilidade na análise de tabelas separáveis, averiguar-se a separabilidade da tabela.

8.1 Zeros amostrais

Os zeros de amostragem correspondem a realizações com probabilidade positiva de ocorrência mas que não foram registadas na amostra seleccionada. Tal facto está relacionado, por um lado, com a variabilidade decorrente do processo de amostragem e, por outro lado, com a baixa probabilidade de ocorrência de algumas classificações (Bishop, Fienberg and Holland, 1980, página 177).

[1]Note-se que a definição de separabilidade de uma tabela de contingência observada não distingue zeros estruturais de zeros amostrais (Bishop, Fienberg and Holland, 1980, página 182).

Este tipo de zeros acarreta três tipos de problemas: *(1)* Põem em causa os resultados assimptóticos; *(2)* As estimativas de máxima verosimilhança dos parâmetros poderão não existir. Num modelo log-linear é especificado um valor $\log e_i$ para cada cela i, estando implícito que $\log e_i < \infty$. Quando uma cela tem frequência nula, a probabilidade de ocorrência de uma observação nessa cela é zero e portanto também o valor estimado da frequência esperada tem que ser nula ($\hat{e}_i = 0$). Nestes casos, $\log \hat{e}_i$ não está definido e as estimativas de máxima verosimilhança não existem; *(3)* Do ponto de vista prático, é importante identificarem-se este tipo de celas, porque embora em muitos casos não existam EMV, muitos packages estatísticos fornecem valores para estas estimativas (Christensen, 1990, página 286).

Para a resolução deste tipo de situações, várias alternativas têm sido sugeridas.

Um das hipóteses sugeridas na literatura para a resolução do problema, por ventura a mais óbvia, é o aumento da dimensão da amostra. Na base desta prescrição está o facto de que, em teoria, uma amostra suficientemente grande faria desaparecer as frequências nulas. No entanto, na prática, a aplicação deste "remédio" é frequentemente impossível e pode não resolver o problema, na medida em que os zeros pode ainda assim subsistir.

Outra alternativa consiste na adição de um pequeno valor a todas as frequências das celas antes de ser efectuada a análise, como *e.g.*, o valor 0.5 recomendado por Goodman (1971) para modelos saturados. Para modelos não-saturados, este procedimento em geral "alisa" os dados em demasia. De acordo com Agresti (1990; página 249) o efeito da adição de uma constante a todas as celas, ou às celas com frequência nula, pode mesmo ser devastador. De facto, o efeito de "alisamento" dos dados encaminha a sua modelação para a simples equiprobabilidade das celas e resulta num procedimento conservativo que tenderá a subestimar o efeito dos parâmetros e a sua significância. Na adopção desta solução, Agresti (1990; página 250) recomenda a realização de uma análise de sensibilidade.

Christensen (1990, página 289) advoga que, no caso de existirem zeros de amostragem, deve-se (1) identificar todas as celas que implicam que as EMV não existam, isto é, identificar as celas para as quais as equações de máxima verosimilhança implicam estimativas de frequências esperadas nulas; (2) remover essas celas do modelo e (3) determinar as EMV para as restantes celas. Na prática, este procedimento equivale a tratar as celas com estimativas de frequências esperadas nulas como zeros estruturais. Este procedimento assume as probabilidades associadas às celas de estimativas nulas como parâmetros perturbadores (*nuisance parameters*), na medida em que correspondem a classificações raras (de reduzida probabilidade) (Aoki and Takemura, 2005). Neste caso, a utilização do algoritmo de Newton-Raphson na estimação das frequências esperadas não oferece problemas, na medida em que este algoritmo ultrapassa sem dificuldade o facto nem todas as combinações das categorias das variáveis serem consideradas no modelo (Christensen, 1990, página 280). Já o uso do método de ajustamento proporcional iterativo pressupõe

a existência de todas as combinações das categorias das variáveis. Neste caso, o inicio do processo iterativo requer que se tome como estimativas iniciais para as celas não nulas, o valor 1 e para as celas nulas, o valor 0 (Christensen, 1990, página 280).

Nesta abordagem, o número de graus de liberdade para a tabela é dado pela diferença entre o número total de celas e o número de zeros. Os graus de liberdade do modelo são determinados como usualmente sendo-lhe subtraído o número de graus de liberdade perdidos por não haver informação disponível para incluir no modelo alguns parâmetros (Christensen, 1990, página 289). No caso dos modelos quase-log-lineares ordinais, o número de graus de liberdade deve ainda ter em conta o número de parâmetros adicionados ao modelo por forma a considerar a ordinalidade das variáveis.

8.2 Zeros estruturais

8.2.1 Tabelas bidimensionais

Tal como visto anteriormente, numa tabela bidimensional completa, o modelo de independência baseia-se na independência estatística entre duas variáveis definida por

$$\pi_{ij} = \pi_{i\bullet} \times \pi_{\bullet j}$$

Em 1968, Goodman sugere que a extensão natural desta definição a tabelas incompletas é

$$\pi_{ij} = \pi_{i\bullet} \times \pi_{\bullet j} \qquad \text{para todas as celas não nulas}$$

Esta relação define a *quase-independência* entre duas variáveis. Assim, definindo S como o sub-conjunto de celas não nulas de uma tabela bidimensional incompleta, pode formular-se o *Modelo Log-linear de Quase-independência* da seguinte forma (Bishop, Fienberg and Holland, 1980, página 179)

$$\log e_{ij} = \mu + \lambda_i^X + \lambda_j^Y \quad \text{para } (i,j) \in S \tag{8.1}$$

Em tabelas bidimensionais incompletas, pode ainda formular-se o *Modelo Quase-saturado* (Bishop, Fienberg and Holland, 1980, página 179)

$$\log e_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad \text{para } (i,j) \in S \tag{8.2}$$

com

$$\sum_{i=1}^{l} \delta_i^{(Y)} \lambda_i^X = \sum_{j=1}^{c} \delta_j^{(X)} \lambda_j^Y = 0 \tag{8.3}$$

$$\sum_{i=1}^{l} \delta_{ij} \lambda_{ij}^{XY} = \sum_{j=1}^{c} \delta_{ij} \lambda_{ij}^{XY} = 0 \tag{8.4}$$

onde

$$\delta_{ij} = \begin{cases} 1 & \text{para } (i,j) \in S \\ 0 & \text{c.c.} \end{cases} \tag{8.5}$$

$$\delta_i^{(Y)} = \begin{cases} 1 & \text{se } \delta_{ij} = 1 \text{ para algum } j \\ 0 & \text{c.c.} \end{cases} \tag{8.6}$$

$$\delta_j^{(X)} = \begin{cases} 1 & \text{se } \delta_{ij} = 1 \text{ para algum } i \\ 0 & \text{c.c.} \end{cases} \tag{8.7}$$

Adaptando estes modelos à presença de variáveis ordinais obtém-se

$$\log e_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta(u_i - \bar{u})(v_j - \bar{v}) \quad \text{para } (i,j) \in S \tag{8.8}$$

com

$$\sum_{i=1}^{l} \delta_i^{(Y)} \lambda_i^X = \sum_{j=1}^{c} \delta_j^{(X)} \lambda_j^Y = 0 \tag{8.9}$$

$$\sum_{i=1}^{l} \delta_{ij}(u_i - \bar{u})(v_j - \bar{v}) = \sum_{j=1}^{c} \delta_{ij}(u_i - \bar{u})(v_j - \bar{v}) = 0 \tag{8.10}$$

onde δ_{ij}, $\delta_i^{(2)}$ e $\delta_j^{(1)}$ são definidos como em (8.5), (8.6) e (8.7).

Por analogia, designaremos estes modelos por *Modelos Quase-log-lineares Ordinais*.

A adaptação destes modelos à presença de zeros amostrais segundo a abordagem de Christensen (1990) implica apenas que a definição dos modelos seja feita para como o sub-conjunto de celas com frequências esperadas não nulas, isto é, o sub-conjunto, designado doravante por S', onde $\hat{e}_{ij} = 0$ para $(i,j,k) \notin S'$.

8.2.2 Tabelas tridimensionais

Os modelos quase-log-lineares para tabelas multidimensionais não são mais que uma generalização dos modelos descritos na subsecção anterior.

Considere-se, pois, o sub-conjunto S de celas sem zeros estruturais de uma tabela $l \times c \times s$ incompleta e seja e_{ijk} as frequências esperadas na cela (i,j,k) onde $e_{ijk} = 0$ para $(i,j,k) \notin S$.

O modelo quase-saturado em tabelas tridimensionais definido para as celas contidas em S é (Bishop, Fienberg and Holland, 1980, secção 5.4.1)

$$\log e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ} \quad \text{para } (i,j,k) \in S \tag{8.11}$$

com, *e.g.*,

$$\sum_{i=1}^{l} \delta_i^{(YZ)} \lambda_i^X = \sum_{i=1}^{l} \delta_{ij}^{(Z)} \lambda_{ij}^{XY} = \sum_{i=1}^{l} \delta_{ik}^{(Y)} \lambda_{ik}^{XZ} = \sum_{i=1}^{l} \delta_{ijk} \lambda_{ijk}^{XYZ} = 0 \tag{8.12}$$

onde

$$\delta_{ijk} = \begin{cases} 1 & \text{para } (i,j,k) \in S \\ 0 & \text{c.c.} \end{cases} \tag{8.13}$$

$$\delta_{ij}^{(Z)} = \begin{cases} 1 & \text{se } \delta_{ijk} = 1 \text{ para algum } k \\ 0 & \text{c.c.} \end{cases} \tag{8.14}$$

$$\delta_{i}^{(YZ)} = \begin{cases} 1 & \text{se } \delta_{ijk} = 1 \text{ para algum } (j,k) \\ 0 & \text{c.c.} \end{cases} \tag{8.15}$$

sendo $\delta_{ik}^{(Y)}$, $\delta_{jk}^{(X)}$, $\delta_{j}^{(XZ)}$ e $\delta_{k}^{(XY)}$ definidos analogamente.

Em tabelas tridimensionais com variáveis ordinais, a adaptação à presença de zeros faz-se analogamente, definindo os modelos log-lineares ordinais no sub-conjunto S' de celas com estimativas não nulas.

Capítulo 9

Analise da associação entre a constituição nutricional de crianças e pais

A segunda parte deste trabalho tem por objectivo apresentar um caso prático de aplicação dos modelos quase-log-lineares ordinais, cujos fundamentos teóricos se desenvolveram nos capítulos anteriores.

Os métodos estatísticos usados na análise dos dados foram implementados usando a linguagem *R for Windows* versão 2.4.0 (http://cran.r-project.org) (ver *script* no CD em anexo).

9.1 Descrição do problema

Hoje em dia, a prevenção da obesidade assim como o seu diagnóstico e tratamento precoces são metas fundamentais para qualquer programa de saúde, sobretudo em pediatria (Melo, A., comunicação pessoal).

O conhecimento dos factores de risco associados ao excesso de peso e obesidade nas crianças tem uma particular importância na medida em que permite uma intervenção precoce e preventiva desta patologia. Vários estudos referem que a obesidade tende a residir dentro das famílias, "passando" de pais para filhos, em resultado da influência da genética, de factores ambientais e da sua interacção (Maes, Neale and Eaves, 1997).

O prévio conhecimento da probabilidade de uma criança possuir excesso de peso ou de ser obesa, num determinado contexto familiar, pode pois constituir uma ferramenta fundamental na avaliação clínica do risco de obesidade e na intervenção clínica precoce.

Este estudo, incluído num projecto mais amplo na área da obesidade infantil, tem como objectivo analisar e caracterizar padrões de associação entre as características nutricionais

das crianças e as características nutricionais dos respectivos pais.

9.2 Selecção da amostra e delineamento experimental

Entre Fevereiro de 2005 e Julho de 2006, foram amostradas 586 crianças de ambos os sexos, seguidas no programa de Saúde Escolar do Centro de Saúde do Lumiar em Lisboa, pertencentes às escolas do 1º ciclo do ensino básico, com idades compreendidas entre os 6 e os 10 anos.

Naturalmente, a população objectivo do estudo pretende-se mais ampla do que a referida, havendo a pretensão da extensão das conclusões à população de crianças pertencentes à faixa etária 6 a 10 anos, (pelo menos) no município de Lisboa. Esta extensão será legitima desde quando a ligação probabilística entre a amostra e a população não se altere com tal expansão (Paulino e Singer, 2006; página 10).

A caracterização clínica da constituição nutricional de indivíduos é tipicamente feita em categorias ordinais resultantes de uma classificação baseada no índice de massa corporal. A classificação das crianças em função da sua categoria nutricional e da constituição nutricional dos respectivos pais consta da Tabela 9.1.

Neste texto, as variáveis serão representadas por M (constituição nutricional da mãe, com i=1,...,6 categorias), P (constituição nutricional do pai, com j=1,...,6 categorias) e C (constituição nutricional da criança, com k=1,2,3 categorias) (ver tabela 9.1).

9.3 Modelo probabilístico

Do exposto na secção anterior, decorre que a estratégia adoptada na selecção das crianças foi a de recolher informação relativamente a tantas crianças quanto possível num determinado período de tempo. Admitindo que todas as variáveis aleatórias associadas às frequências o_{ijk}, $(i = 1, ..., 6; j = 1, ..., 6; k = 1, 2, 3)$ são independentes e que em cada uma das celas da tabela:

1. o número de crianças observadas num determinado intervalo de tempo é independente do número de crianças observadas em qualquer outro intervalo de tempo disjunto do primeiro;

2. o processo de chegadas de crianças é *homogéneo no tempo*, isto é, a distribuição do número de crianças só depende do comprimento do intervalo de tempo considerado, não importando quando tem início;

3. a probabilidade de chegada de uma criança cresce regular e linearmente com a amplitude do intervalo;

Tabela 9.1: Dados sobre a constituição nutricional de crianças e pais

		Classe nutricional criança		
Classe nutricional mãe	**Classe nutricional pai**	*Eutrófico*	*Excesso de Peso*	*Obesidade*
Baixo Peso	*Baixo Peso*	2	0	0
	Eutrófico	18	2	1
	Excesso de Peso	16	2	4
	Obesidade Tipo I	3	1	0
	Obesidade Tipo II	0	0	0
	Obesidade Tipo III	0	0	0
Eutrófico	*Baixo Peso*	5	0	0
	Eutrófico	104	15	13
	Excesso de Peso	95	31	33
	Obesidade Tipo I	16	2	9
	Obesidade Tipo II	1	0	1
	Obesidade Tipo III	1	0	1
Excesso de Peso	*Baixo Peso*	2	0	0
	Eutrófico	30	7	5
	Excesso de Peso	42	24	26
	Obesidade Tipo I	11	0	10
	Obesidade Tipo II	2	3	0
	Obesidade Tipo III	0	0	1
Obesidade Tipo I	*Baixo Peso*	0	0	0
	Eutrófico	6	0	1
	Excesso de Peso	10	6	4
	Obesidade Tipo I	3	1	5
	Obesidade Tipo II	0	0	0
	Obesidade Tipo III	0	0	0
Obesidade Tipo II	*Baixo Peso*	0	0	0
	Eutrófico	1	0	0
	Excesso de Peso	3	0	1
	Obesidade Tipo I	0	0	1
	Obesidade Tipo II	0	0	0
	Obesidade Tipo III	0	0	0
Obesidade Tipo III	*Baixo Peso*	0	0	0
	Eutrófico	1	2	1
	Excesso de Peso	0	0	0
	Obesidade Tipo I	0	0	1
	Obesidade Tipo II	0	0	0
	Obesidade Tipo III	0	0	0

4. a probabilidade de chegadas simultâneas em intervalos de tempo suficientemente pequenos é desprezável,

então, o modelo probabilístico adequado à modelação das ocorrências é o Produto de Distribuições Poisson (equação (3.3) para $i = 1, ..., 6; j = 1, ..., 6; k = 1, 2, 3$).

9.4 Análise dos zeros amostrais

A tabela 9.1 evidencia a presença de múltiplos zeros amostrais. A primeira preocupação na análise dos dados consistiu precisamente em delinear o tratamento destas observações.

Nos dados em causa os zeros amostrais ocorrem em celas correspondente a classificações de reduzida probabilidade, pelo que o seu tratamento foi feito, segundo a abordagem de Christensen (1990), como se de zeros estruturais se tratasse. Na prática, a presença destes zeros teve duas consequências em termos de análise log-linear:

- As celas para as quais as equações de máxima verosimilhança implicavam estimativas de frequências esperadas nulas foram identificadas e removidas do modelo. As estimativas de máxima verosimilhança foram encontradas para as restantes celas;

- O nº de graus de liberdade para a tabela e para os termos dos modelos foi ajustado em função do numero de zeros "fixos". O nº de graus de liberdade para um modelo foi obtido pela diferença entre o nº de graus de g.l. da tabela completa e o nº de g.l. perdidos por não haver informação disponível relativamente a alguns parâmetros.

9.4.1 Identificação das celas com frequências estimadas nulas

Analisando a tabela 9.1 torna-se evidente que o modelo saturado não pode ser ajustado, já que, neste modelo, $\hat{e}_{ijk} = o_{ijk}$. Porque o modelo a ajustar é log-linear, o $\log \hat{e}_{ijk}$ tem de estar definido. Ora, nas celas onde $o_{ijk} = 0$ ou $\hat{e}_{ijk} \neq o_{ijk}$ (o que não pode acontecer num modelo saturado) ou $\log \hat{e}_{ijk}$ não está definido.

Considerando o modelo sem interacção de 2ª ordem $[MP][MC][PC]$, têm-se as equações de verosimilhança, para todo o i, j e k

$$\hat{e}_{ij\bullet} = o_{ij\bullet} \qquad \hat{e}_{i\bullet k} = o_{i\bullet k} \qquad \hat{e}_{\bullet jk} = o_{\bullet jk}$$

Nas tabelas marginais têm-se vários totais nulos. Assim, na tabela [MP] (Tabela 9.2) há 12 zeros: $o_{15\bullet} = o_{16\bullet} = o_{41\bullet} = o_{45\bullet} = o_{46\bullet} = o_{51\bullet} = o_{55\bullet} = o_{56\bullet} = o_{61\bullet} = o_{63\bullet} = o_{65\bullet} = o_{66\bullet} = 0$.

Este totais nulos envolvem 36 celas de frequência nula

$$
\begin{array}{cccccc}
(1,5,1) & (1,5,2) & (1,5,3) & (1,6,1) & (1,6,2) & (1,6,3) \\
(4,1,1) & (4,1,2) & (4,1,3) & (4,5,1) & (4,5,2) & (4,5,3) \\
(4,6,1) & (4,6,2) & (4,6,3) & (5,1,1) & (5,1,2) & (5,1,3) \\
(5,5,1) & (5,5,2) & (5,5,3) & (5,6,1) & (5,6,2) & (5,6,3) \\
(6,1,1) & (6,1,2) & (6,1,3) & (6,3,1) & (6,3,2) & (6,3,3) \\
(6,5,1) & (6,5,2) & (6,5,3) & (6,6,1) & (6,6,2) & (6,6,3)
\end{array}
$$

A tabela marginal [MC] (Tabela 9.3) tem 1 zero: $o_{5\bullet 2} = 0$, que envolve 6 celas de frequência nula

$$
\begin{array}{cccccc}
(5,1,2) & (5,2,2) & (5,3,2) & (5,4,2) & (5,5,2) & (5,6,2)
\end{array}
$$

Por último, a tabela marginal $[PC]$ (Tabela 9.4) tem três zeros: $o_{\bullet 12} = o_{\bullet 13} = o_{\bullet 62} = 0$, envolvendo 18 celas de frequência nula

$$
\begin{array}{cccccc}
(1,1,2) & (2,1,2) & (3,1,2) & (4,1,2) & (5,1,2) & (6,1,2) \\
(1,1,3) & (2,1,3) & (3,1,3) & (4,1,3) & (5,1,3) & (6,1,3) \\
(1,6,2) & (2,6,2) & (3,6,2) & (4,6,2) & (5,6,2) & (6,6,3)
\end{array}
$$

Todas as celas listadas possuem $\hat{e} = 0$, havendo 47 celas distintas nesta situação.

9.4.2 Cálculo do número de graus de liberdade

Recorde-se que a tabela 9.1 possui um total de 108 celas. Assim, atendendo às 47 celas distintas para as quais se estimam frequências esperadas nulas, o número de g.l. da tabela é $108 - 47 = 61$.

Considere-se agora os g.l. dos termos dos modelos. A tabela marginal [MP] (tabela 9.2) possui 12 zeros (*i.e.*, $o_{ij\bullet} = 0$). Assim, o número de g.l. da tabela 6×6 não será 36, mas $36 - 12 = 24$, sendo a afectação destes g.l. pelos termos do modelo a seguinte

$$\mu : 1 \qquad M : 5 \qquad P : 5 \qquad MP : 13\ (5 \times 5 - 12)$$

A tabela marginal [MC] (tabela 9.3) com 1 zero, possui $6 \times 3 - 1 = 17$ g.l. afectos aos termos do modelo do seguinte modo

$$\mu : 1 \qquad M : 5 \qquad C : 2 \qquad MC : 9\ (5 \times 2 - 1)$$

Tabela 9.2: Tabela marginal [MP]

	P					
M	**1**	**2**	**3**	**4**	**5**	**6**
1	2	5	2	0	0	0
2	21	132	42	7	1	4
3	22	159	92	20	4	0
4	4	27	21	9	1	1
5	0	2	5	0	0	0
6	0	2	1	0	0	0

Tabela 9.3: Tabela marginal [MC]

	C		
M	**1**	**2**	**3**
1	39	5	5
2	222	48	57
3	87	34	42
4	19	7	10
5	4	0	2
6	1	2	2

Tabela 9.4: Tabela marginal [PC]

	C		
P	**1**	**2**	**3**
1	3	0	0
2	160	26	21
3	166	63	68
4	33	4	26
5	3	3	1
6	1	0	2

De forma análoga, a tabela marginal [PC] (tabela 9.4) com 3 zeros, possui $6\times3-3=15$ g.l. com a seguinte distribuição pelos termos do modelo

$$\mu:1 \qquad P:5 \qquad C:2 \qquad PC:7\ (5\times2-3)$$

Recorde-se que a tabela $6\times6\times3$ (Tabela 9.1) possui 61 g.l.. Assim, o nº de parâmetros independentes associado ao termos representativo da interacção de 2ª ordem será $61-(1+5+5+2+13+9+7)=19$. Na tabela 9.5, resume-se o nº de parâmetros independentes associados aos termos dos modelos para a tabela 9.1.

Tabela 9.5: Número de parâmetros independentes associados aos termos dos modelos para a tabela 9.1

Termo	*Tabela completa*	*Tabela incompleta*
μ	1	1
λ_i^M	5	5
λ_j^P	5	5
λ_k^C	2	2
λ_{ij}^{MP}	25	13
λ_{ik}^{MC}	10	9
λ_{jk}^{PC}	10	7
λ_{ijk}^{MPC}	50	19

A determinação do número de parâmetros independentes permitiu o cálculo do número de g.l. associados aos modelos[1], através do método 2 exposto na secção 6.1, obtendo-se

[1] A notação adoptada para designar os modelos quase-log-lineares será a mesma adoptada para os modelos em tabelas completas acrescida do condicionamento a S'. Assim, por exemplo, o modelo de independência em tabelas tridimensionais será denotado por $[X][Y][Z]\big|_{S'}$.

os seguintes resultados

$$
\begin{aligned}
&[MPC]|_{S'} : 0 \\
&[MP][MC][PC]|_{S} : 19 \\
&[MP][MC]|_{S'} : 19 + 7 = 26 \\
&[MP][PC]|_{S'} : 19 + 9 = 28 \\
&[MC][PC]|_{S'} : 19 + 13 = 32 \\
&[PC][M]|_{S'} : 19 + 9 + 13 = 41 \\
&[MC][P]|_{S'} : 19 + 7 + 13 = 39 \\
&[MP][C]|_{S'} : 19 + 7 + 9 = 35 \\
&[M][P][C]|_{S'} : 19 + 7 + 9 + 13 = 48
\end{aligned}
$$

9.5 Ajustamento de modelos quase-log-lineares

Para proceder ao ajustamento de modelos quase-log-lineares as 47 celas de frequência estimada nula foram tratadas como celas de zeros estruturais e portanto removidas do modelo. As estimativas de máxima verosimilhança foram calculadas apenas para as restantes celas da tabela. Nestas condições foi definido o sub-conjunto S' de celas onde $\hat{e}_{ijk} \neq 0$. Para tal, a tabela dos dados foi aumentada com uma variável indicatriz, δ_{ijk}, distinguido as celas com estimativas não nulas das celas com estimativas nulas.

$$
\delta_{ijk} = \begin{cases} 1 & (i,j,k) \in S' \\ 0 & (i,j,k) \notin S' \end{cases}
$$

O ajustamento dos modelos foi feito usando a função `glm`, fixando o argumento `family = poisson`, sendo a estimação feita pelo Método de Newton-Raphson.

9.5.1 Comparação dos modelos quase-log-lineares

Tal como salientado anteriormente, o facto de todas as variáveis em análise serem ordinais aumenta muito o número total de modelos log-lineares possíveis de ajustar. Assim, para seleccionar o modelo quase-log-linear ordinal que melhor se ajusta aos dados procedeu-se de acordo com o exposto na sub-secção 7.3.

Sob o modelo probabilístico (3.3), analisou-se, numa primeira fase, o ajustamento de todos os modelos quase-log-lineares (não ordinais).

Com base nas estatísticas (6.1), (6.2) e (6.14) obtiveram-se os resultados de ajustamento global apresentados na tabela 9.6.

Os valores-p associados às estatísticas X^2 e Y^2 revelam que os modelos $[MP][MC][PC]|_{S'}$, $[MP][PC]|_{S'}$ e $[MC](PC]|_{S'}$ são todos elegíveis, para um nível de

Tabela 9.6: Ajustamento dos modelos quase-log-lineares

Modelo	*g.l.*	Y^2	p	X^2	p	A
$[MPC]\|_{S'}$	0	0.0000	–	0.0000	–	0
$[MP][MC][PC]\|_{S'}$	19	19.9727	0.396	17.0316	0.588	−18.027
$[MP][MC]\|_{S'}$	26	66.4407	0.000	60.2371	0.000	14.440
$[MP][PC]\|_{S'}$	28	37.1497	0.116	33.5651	0.216	−18.850
$[MC][PC]\|_{S'}$	32	41.0218	0.132	37.6643	0.226	−22.978
$[PC][M]\|_{S'}$	41	64.6015	0.011	63.1281	0.015	−17.399
$[MC][P]\|_{S'}$	39	93.1660	0.000	96.6697	0.000	15.166
$[MP][C]\|_{S'}$	35	87.5058	0.000	86.1583	0.000	17.506
$[M][P][C]\|_{S'}$	48	114.8034	0.000	136.6005	0.000	18.803

significância de 0.05. Os restantes modelos possuem ajustamentos muito pobres às frequências observadas.

Para decidir sobre a inclusão dos termos representativos das interacções de 1ª ordem considere-se os resultados da comparação dos modelos encaixados resumidos na tabela 9.7. Considerando o nível de significância 0.05, rejeitam-se as hipóteses H_0^1 e H_0^2 e não se rejeita H_0^3. Tal significa que o modelo adequado para descrever os dados é o modelo $[MC][PC]|_{S'}$.

Tabela 9.7: Comparação dos modelos quase-log-lineares encaixados sem ordinalidade

Hipóteses	M_a	M_b	$Y^2(M_b\|M_a)$	$g.l.$	p
$H_0^1 : \lambda_{jk}^{PC} = 0$	$[MP][MC]\|_{S'}$	$[MP][MC][PC]\|_{S'}$	46.4680	7	0.000
$H_0^2 : \lambda_{ik}^{MC} = 0$	$[MP][PC]\|_{S'}$	$[MP][MC][PC]\|_{S'}$	17.1770	9	0.046
$H_0^3 : \lambda_{ij}^{MP} = 0$	$[MC][PC]\|_{S'}$	$[MP][MC][PC]\|_{S'}$	21.0491	13	0.072

Repare-se que a estatística A (tabela 9.6) aponta claramente para a eleição do modelo $[MC][PC]|_{S'}$ assumindo para este um valor mínimo ($A([MC][PC]|_{S'}) = -22.978$). A utilização da medida de Akaike revela-se assim mais simples e mais eficaz na comparação de modelos.

9.6 Ajustamento de modelos quase-log-lineares ordinais

Para considerar a ordinalidade das variáveis foram usados os *scores* com distâncias unitárias definidos pelos próprios índices das categorias ordenadas crescentemente, pelas razões expostas no capítulo 7.

Tendo como guia os resultados obtidos na secção anterior, foi, numa primeira fase, ajustado aos dados o modelo de quase-associação uniforme heterogénea

$$\begin{aligned}\log e_{ijk} = \mu + \lambda_i^M + \lambda_j^P + \lambda_k^C + \beta^{MP}(u_i - \bar{u})(v_j - \bar{v}) + \beta^{MC}(u_i - \bar{u})(w_k - \bar{w}) \\ +\beta^{PC}(v_j - \bar{v})(w_k - \bar{w}) + \beta^{MPC}(u_i - \bar{u})(v_j - \bar{v})(w_k - \bar{w}) \quad \text{para } (i,j,k) \in S' \end{aligned} \tag{9.1}$$

e, em seguida, ajustados os modelos de quase-associação uniforme homogénea (secção 6.2)

$$\begin{aligned}\log e_{ijk} = \mu + \lambda_i^M + \lambda_j^P + \lambda_k^C + \beta^{MP}(u_i - \bar{u})(v_j - \bar{v}) + \beta^{MC}(u_i - \bar{u})(w_k - \bar{w}) \\ +\beta^{PC}(v_j - \bar{v})(w_k - \bar{w}) \quad \text{para } (i,j,k) \in S' \end{aligned} \tag{9.2}$$

$$\begin{aligned}\log e_{ijk} = \mu + \lambda_i^M + \lambda_j^P + \lambda_k^C + \beta^{MP}(u_i - \bar{u})(v_j - \bar{v}) + \beta^{MC}(u_i - \bar{u})(w_k - \bar{w}) \\ \text{para } (i,j,k) \in S' \end{aligned} \tag{9.3}$$

$$\begin{aligned}\log e_{ijk} = \mu + \lambda_i^M + \lambda_j^P + \lambda_k^C + \beta^{MP}(u_i - \bar{u})(v_j - \bar{v}) + \beta^{PC}(u_i - \bar{u})(w_k - \bar{w}) \\ \text{para } (i,j,k) \in S' \end{aligned} \tag{9.4}$$

$$\begin{aligned}\log e_{ijk} = \mu + \lambda_i^M + \lambda_j^P + \lambda_k^C + \beta^{MC}(u_i - \bar{u})(v_j - \bar{v}) + \beta^{PC}(u_i - \bar{u})(w_k - \bar{w}) \\ \text{para } (i,j,k) \in S' \end{aligned} \tag{9.5}$$

9.6.1 Comparação dos modelos quase-log-lineares ordinais

Os resultados relativos à comparação dos modelos (9.1) a (9.5) encontram-se resumidos na tabela 9.8.

Os valores da medida de Akaike levam a concluir que o modelo $[\beta^{MP}][\beta^{MC}][\beta^{PC}]|_{S'}$ (equação 9.2) é o que melhor descreve os dados. Repare-se como os valores desta medida

Tabela 9.8: Medida de informação de Akaike para os modelos quase-log-lineares ordinais

Modelo	A
$[\beta^{MP}][\beta^{MC}][\beta^{PC}]\|_{S'}$	-31.255
$[\beta^{MP}][\beta^{PC}]\|_{S'}$	-24.463
$[\beta^{MC}][\beta^{PC}]\|_{S'}$	-25.980
$[\beta^{MP}][\beta^{MC}]\|_{S'}$	-8.260

são notoriamente inferiores aos obtidos para os correspondentes modelos quase-log-lineares antes da consideração da ordinalidade das variáveis (tabela 9.6), denotando uma clara melhoria no ajustamento.

Os resultados apresentados na tabela 9.9 dizem respeito aos testes de ajustamento correspondentes à comparação múltipla de três modelos encaixados.

Tabela 9.9: Comparação dos modelos quase-log-lineares ordinais encaixados

Hipóteses	M_a	M_b	$Y^2(M_a\|M_b)$	$g.l.$	p
$H_0^1 : \beta^{MPC} = 0$	$[\beta^{MP}][\beta^{MC}][\beta^{PC}]\|_{S'}$	$[\beta^{MPC}]\|_{S'}$	0.0775	1	0.7807
$H_0^2 : \beta^{MC} = 0$	$[\beta^{MP}][\beta^{PC}]\|_{S'}$	$[\beta^{MP}][\beta^{MC}][\beta^{PC}]\|_{S'}$	10.7151	1	0.0011
$H_0^3 : \beta^{MP} = 0$	$[\beta^{MC}][\beta^{PC}]\|_{S'}$	$[\beta^{MP}][\beta^{MC}][\beta^{PC}]\|_{S'}$	9.1985	1	0.0024
$H_0^4 : \beta^{PC} = 0$	$[\beta^{MP}][\beta^{MC}]\|_{S'}$	$[\beta^{MP}][\beta^{MC}][\beta^{PC}]\|_{S'}$	26.9185	1	< 0.001

Para que a probabilidade tipo I não ultrapasse $\alpha = 0.05$ na globalidade dos testes efectuados, cada um dos testes foi efectuado a um nível de significância de $1 - \sqrt[4]{1 - 0.05} = 0.013$. Considerando este nível de significância, os testes indicam que o termo representativo da interacção de 2ª ordem não é significativamente diferente de zero (p = 0.7807) sendo portanto removido do modelo. Por outro lado, ao contrário do que aconteceu quando as variáveis foram consideradas como nominais, todos os termos relativos às interacções de primeira ordem são significativos (p=0.0011; p= 0.0024 e p<0.0010) permanecendo, por isso, no modelo. Repare-se que apenas a consideração da ordinalidade das variáveis permitiu evidenciar a associação M-P. Esta análise conduz também à escolha do modelo $[\beta^{MP}][\beta^{MC}][\beta^{PC}]\|_{S'}$ como o melhor modelo.

9.7 Análise dos resíduos

A tabela 9.10 contem as estimativas das frequências esperadas. Com base nas frequências observadas e nos valores destas estimativas calcularam-se os resíduos ajustados (equação 6.5), cuja representação gráfica é apresentada na Figura 9.1(a). As figuras 9.1(b), (c) e (d) ilustram a distribuição empírica dos resíduos ajustados e a aproximação à Distribuição Normal. Os gráficos mostram uma boa adequação dos resíduos ajustados à distribuição Normal Padrão, havendo apenas um pequeno desvio para valores mais elevados. O teste de ajustamento de Kolmogorov-Smirnov permitiu validar de um modo formal a aderência à Distribuição Normal Padrão ($d = 0.1042; p = 0.4891$). Assim, aplicando a correcção de Bonferroni, um resíduo será significativo quando o seu valor absoluto exceder o valor $z_{1-0.05/(2\times 61)} = z_{0.9996} = 3.35$, para um nível de significância global de 0.05. Tal como se pode ver na figura 9.1(a) nenhum resíduo ajustado excede este valor critico absoluto, pelo que se pode concluir que o modelo possui um bom ajustamento interno. Em nenhuma cela se verifica uma influência nefasta estatisticamente significativa no ajustamento global.

9.8 Análise dos parâmetros

Na tabela 9.11 apresentam-se as estimativas dos parâmetros e respectivos erros padrão do modelo ajustado, usando a parametrização em termos da cela de referência $(1, 1, 1)$ (isto é, $\lambda_1^M = \lambda_1^P = \lambda_1^C = 0$).

A comparação dos valores absolutos padronizados — $S(\hat{\lambda})$— permite estabelecer que a importância das variáveis na determinação das frequências aumenta à medida que se sobe na classe nutricional. Em particular, verifica-se que o número de crianças eutróficas tende a ser estocasticamente menor à medida que aumenta a classe nutricional à qual pertencem a mãe ou o pai, sendo este efeito tanto mais marcado quanto mais alta é a classe nutricional à qual pertencem os pais.

As estimativas dos parâmetros representativos das interacções evidenciam uma associação positiva entre as variáveis, isto é, os membros da família tendem a pertencer a classes nutricionais mais altas/baixas quando os restantes membros da família também pertencem. Comparando os valores padronizados, verifica-se que a classe nutricional da criança está mais fortemente associada à classe nutricional do pai do que à classe nutricional da mãe. Por outro lado, a associação entre as classes nutricionais do pai e da mãe é mais fraca que a associação entre a classe nutricional da criança e respectivos progenitores.

Estes resultados permitem ainda testar as hipóteses unilaterais $H_0 : \beta \leq 0$ vs. $H_1 : \beta > 0$. Na tabela 9.12 apresentam-se os valores-p associados aos testes de significância das interacções de 1ª ordem.

Para qualquer um dos parâmetros, a hipótese nula é rejeitada evidenciando uma signi-

Tabela 9.10: EMV das frequências esperadas

		Classe nutricional criança		
Classe nutricional mãe	**Classe nutricional pai**	*Eutrófico*	*Excesso de Peso*	*Obesidade*
Baixo Peso	*Baixo Peso*	1.4	0.0	0.0
	Eutrófico	17.5	2.7	1.6
	Excesso de Peso	16.0	3.6	3.2
	Obesidade Tipo I	1.8	0.6	0.7
	Obesidade Tipo II	0.0	0.0	0.0
	Obesidade Tipo III	0.0	0.0	0.0
Eutrófico	*Baixo Peso*	5.9	0.0	0.0
	Eutrófico	92.7	17.4	13.0
	Excesso de Peso	104.5	28.6	31.0
	Obesidade Tipo I	14.3	5.7	9.0
	Obesidade Tipo II	1.3	0.7	1.7
	Obesidade Tipo III	0.4	0.0	1.0
Excesso de Peso	*Baixo Peso*	1.8	0.0	0.0
	Eutrófico	34.7	8.0	7.3
	Excesso de Peso	48.4	16.2	21.6
	Obesidade Tipo I	8.2	4.0	7.7
	Obesidade Tipo II	0.9	0.6	1.8
	Obesidade Tipo III	0.3	0.0	1.3
Obesidade Tipo I	*Baixo Peso*	0.0	0.0	0.0
	Eutrófico	5.7	1.6	1.8
	Excesso de Peso	9.9	4.1	6.6
	Obesidade Tipo I	2.1	1.2	2.9
	Obesidade Tipo II	0.0	0.0	0.0
	Obesidade Tipo III	0.0	0.0	0.0
Obesidade Tipo II	*Baixo Peso*	0.0	0.0	0.0
	Eutrófico	0.8	0.0	0.4
	Excesso de Peso	1.7	0.0	1.7
	Obesidade Tipo I	0.4	0.0	0.9
	Obesidade Tipo II	0.0	0.0	0.0
	Obesidade Tipo III	0.0	0.0	0.0
Obesidade Tipo III	*Baixo Peso*	0.0	0.0	0.0
	Eutrófico	0.8	0.3	0.6
	Excesso de Peso	0.0	0.0	0.0
	Obesidade Tipo I	0.7	0.6	2.1
	Obesidade Tipo II	0.0	0.0	0.0
	Obesidade Tipo III	0.0	0.0	0.0

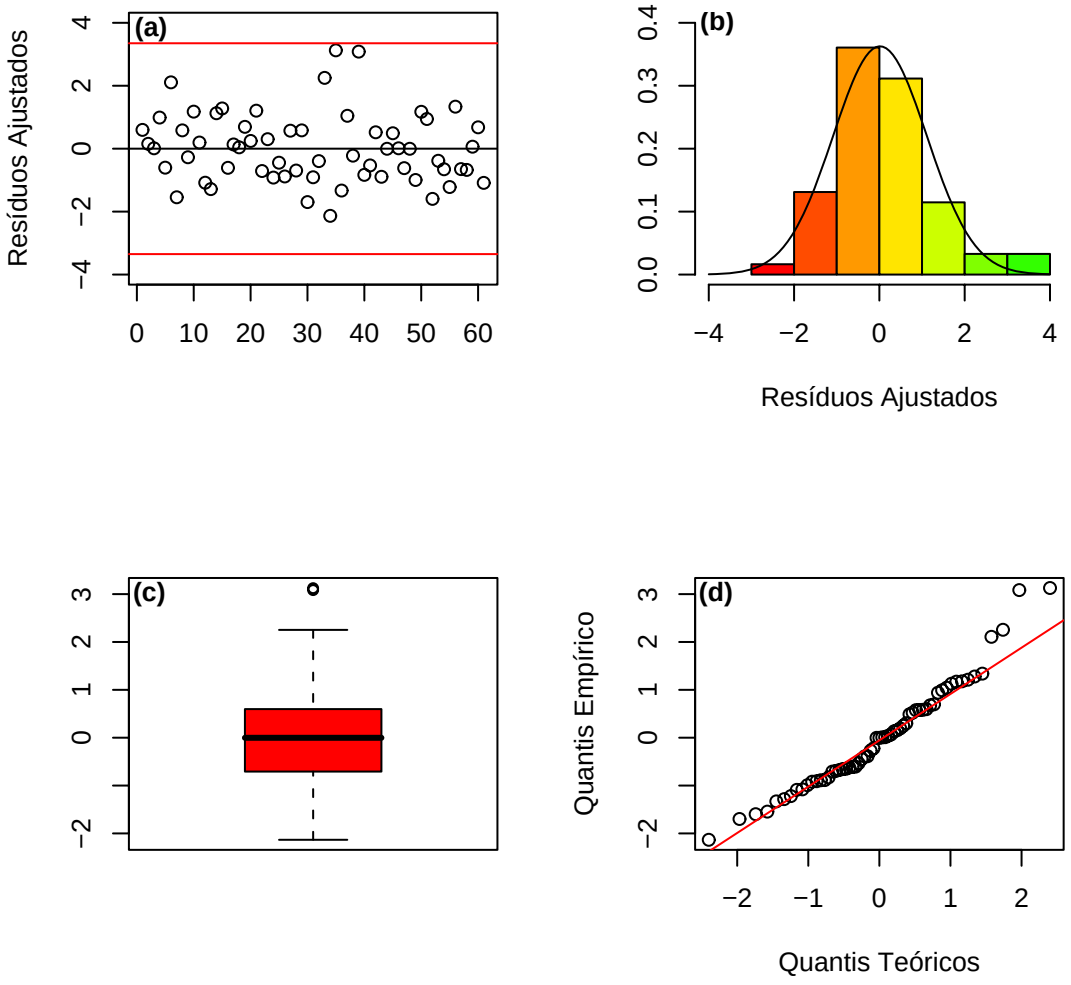

Figura 9.1: Análise gráfica dos resíduos ajustados

Tabela 9.11: Estimativas paramétricas e erros-padrão aproximados no modelo ajustado

Parâmetro	*Estimativa*	*Erro padrão*	$S(\hat{\lambda})$
λ_2^M	1.03540	0.23602	4.387
λ_3^M	−0.57580	0.41538	−1.386
λ_4^M	−3.00688	0.63878	−4.707
λ_5^M	−5.60463	0.94475	−5.932
λ_6^M	−6.26850	1.23205	−5.088
λ_2^P	1.96119	0.36838	5.324
λ_3^P	1.28098	0.46821	2.736
λ_4^P	−1.50677	0.65540	−2.299
λ_5^P	−4.73236	0.92348	−5.124
λ_6^P	−6.79486	1.25059	−5.433
λ_2^C	−2.82543	0.26800	−10.543
λ_3^C	−4.27362	0.52949	−8.071
β^{MP}	0.21307	0.07124	2.991
β^{MC}	0.20346	0.06253	3.254
β^{PC}	0.37409	0.07471	5.007

Tabela 9.12: Resultados dos testes unilaterais aos parâmetros β no modelo final

Hipótese nula	p
$H_0 : \beta^{MP} \leq 0$	0.0028
$H_0 : \beta^{MC} \leq 0$	0.0011
$H_0 : \beta^{PC} \leq 0$	< 0.0001

ficativa associação positiva entre as variáveis. Para um nível fixo de P a razão de chances condicionais locais entre M e C é $\exp(0.2035) = 1.23$. Tal significa que a probabilidade estimada da criança pertencer a uma classe nutricional mais alta do que à imediatamente anterior é 1.2 vezes maior quando a mãe pertence a uma categoria nutricional mais alta do que quando a mãe pertence a uma categoria nutricional inferior. Quando à influência da classe nutricional do pai, a razão de chances condicionais locais estimada $\exp(0.3741) = 1.45$, indica que a probabilidade estimada da criança pertencer a uma classe nutricional mais alta do que à imediatamente anterior é cerca de uma vez e meia maior quando o pai pertence a uma categoria nutricional mais alta do que quando o pai pertence a uma categoria nutricional inferior. Usando as classes extremas da tabela P-C, a probabilidade estimada de uma criança ser obesa em vez de eutrófica é $\exp(0.3741 \times (3-1) \times (4-2)) = 4.5$ vezes maior quando o seu pai é obeso do que quando o seu pai é eutrófico. Procedendo de modo idêntico relativamente à tabela M-C, conclui-se que a probabilidade estimada de uma criança ser obesa em vez de eutrófica é $\exp(0.2035 \times (3-1) \times (4-2)) = 2.3$ vezes maior quando a sua mãe é obesa do que quando esta é eutrófica.

Capítulo 10

Conclusão

A adaptação dos modelos log-lineares a variáveis ordinais tem início com a atribuição de *scores* ordinais às categorias das variáveis. A sua definição deve ser cuidadosa já que afecta a interpretação dos termos do modelo que reflectem a ordinalidade das variáveis. Na prática, a atribuição dos *scores* deve fazer-se de modo a simplificar a interpretação dos modelos.

Uma vez atribuídos os scores ordinais, a construção de modelos log-lineares ordinais tem por base a substituição dos termos representativos das interacções por termos estruturados que estabelecem os desvios relativamente à independência como função linear de cada uma das variáveis, considerando fixos os níveis das restantes variáveis.

Esta substituição dos termos representativos das interações torna os modelos mais parcimoniosos, já que cada termo estruturado (função dos scores ordinais) tem apenas mais um parâmetro que o modelo geral de independência.

Nos modelos log-linear ordinais as equações de máxima verosimilhança que incluem os scores não têm não têm soluções directas, exigindo um método iterativo para a determinação das EMV.

A modelação log-linear de tabelas com zeros, deve começar pela analise da separabilidade da tabela. Em tabelas não separáveis, o tratamento dos zeros pode ter várias prescrições. Neste estudo explorou-se a abordagem de Christensen (1990) de acordo com a qual, o tratamento de zeros consiste em identificar todas as celas que implicam que as EMV não existam, remover essas celas do modelo e determinar as EMV para as restantes celas. Subjacente a esta abordagen está o pressuposto de que as probabilidades associadas às celas onde ocorrem os zeros podem ser encaradas como parâmetros perturbadores (*nuisance parameters*) na medida em que correspondem a classificações raras (de reduzida probabilidade).

Em termos práticos esta abordagem tem duas consequências:

- A determinação das estimativas de máxima verosimilhança é feita apenas para as

celas da tabela com estimativas não nulas, sendo portanto necessário definir uma variável indicatriz, a incluir no modelo, que distinga celas com estimativas não nulas de celas com estimativas nulas

- A segunda consequência diz respeito à correcção do número de graus de liberdade. O nº de graus de liberdade para a tabela e para os termos dos modelos devem ser ajustados em função do numero de zeros "fixos". O nº de graus de liberdade para um modelo será obtido pela diferença entre o nº de graus de g.l. da tabela completa e o nº de g.l. perdidos por não haver informação disponível relativamente a alguns parâmetros. No caso dos modelos quase-log-lineares ordinais, o número de graus de liberdade deve ainda ter em conta o número de termos de interacção substituídos no modelo clássico por forma a considerar a ordinalidade das variáveis.

No que se refere à selecção dos modelos, há a considerar que a introdução de termos que incluem os *scores* ordinais aumenta muito o número de possíveis modelos a ajustar. Existem por isso várias estratégias possíveis para a selecção do melhor modelo, sendo uma das mais usadas ter como guia os modelos log-lineares sem considerar a ordinalidade das variáveis e, uma vez obtido um modelo log-linear bem ajustado aos dados, proceder-se à substituição faseada dos termos de interacção por termos estruturados baseados nos *scores*.

Na comparação de modelos, conclui-se que a estatística avaliadora do ajustamento que melhor se adequa à selecção do melhor modelo quase-log-linear ordinal foi a medida de Akaike, já que fazendo o balanço entre o quão bem o modelo se ajusta aos dados e a complexidade do modelo, esta estatística é comparável entre quaisquer dois modelos, independentemente do número de graus de liberdade dos modelos e do facto de se tratarem ou não de modelos encaixados.

10.1 Caso prático

No que se refere ao estudo com o objectivo de analisar e caracterizar padrões de associação entre as características nutricionais das crianças e as características nutricionais dos respectivos pais, concluiu-se que existe uma associação positiva entre as variáveis, isto é, os membros da família tendem a pertencer a classes nutricionais mais altas/baixas quando os restantes membros da família também pertencem. Verificou-se que a classe nutricional da criança está mais fortemente associada à classe nutricional do pai do que à classe nutricional da mãe. Por outro lado, a associação entre as classes nutricionais do pai e da mãe é mais fraca que a associação entre a classe nutricional da criança e respectivos progenitores.

O estudo permitiu concluir que a probabilidade de uma criança ter problemas de excesso de peso ou obesidade é muito mais elevada quando os respectivos pais pertencem a classes nutricionais altas do que quando os pais são eutróficos.

Capítulo 11

Referências bibliográficas

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, John Wiley & Sons, NY.

Agresti, A. (1990). *Categorical Data Analysis*, John Wiley & Sons, NY.

Aoki, S. and Takemura, A. (1980). Markov chain Monte Carlo exact tests for incomplete two-way contingency tables, *Journal of Statistical Computation and Simulation*, **75**(10): 787-812.

Benedetti, J. K. and Brown, M. B. (1978). Strategies for the selection of loglinear models. *Biometrics* **34**: 680-686.

Birch, M. W. (1964). Maximum likelihood in three-way contingency tables, *Journal of the Royal Statistical Society, B* **25**: 220-233.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1980). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press, Cambridge.

Brown, M. B. (1976). Screening effects in multidimensional contingency tables. *Applied Statistics* **25**: 37-46.

Casella, G. and Berger, R. (2002). *Statistical Inference*, 2nd edition, Duxbury, Pacific Grove.

Christensen, R. (1990). *Log-linear Models and Logistic Regression*, 2nd Edition, Springer-Verlag, NY.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the*

Royal Statistical Society **46**: 440-464.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annuals of Mathematics and Statistics* **11**: 427-444.

Everitt, B. S. (1994). *The analysis of contingency tables*, Chapman & Hall, London.

Goodman, L. A. (1968). The analysis of cross-classified data: independence, quasi-independence and interaction in contingency tables with or without missing cells., *Journal of the American Statistical Association* **63**:1091-1131.

Goodman, L. A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications., *Journal of the American Statistical Association* **65**:226-256.

Goodman, L. A. (1971). The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for bulding models for multiple classifications, *Technometrics* **13**:33-61.

Maes, H.H., Neale, M.C. and Eaves,L.J. (1997). Genetic and environmental factors in body weight and human adiposity, *Behavioral Genetics*, 27: 325-351.

Paulino, C. D. e Singer, J. M. (2006). *Análise de dados categorizados*, Editora Edgard Blücher, São Paulo.

Pestana, D. e Velosa, S. (2006). *Probabilidade e Estatística*, 2ª edição, Fundação Calouste Gulbenkian, Lisboa.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*, Springer-Verlag, NY.

Upton, G. J. G. (1980). *The analysis of cross-tabulated data*, John Wiley & Sons, NY.