GET00211 - Modelos Lineares 2

Rafael Erbisti

Instituto de Matemática e Estatística

Universidade Federal Fluminense

Aula 14

Exemplos de Tabelas de Contingência

- Antes de especificar um modelo log-linear para dados de frequência resumidos em tabelas de contingência, é importante considerar como o desenho do estudo pode determinar restrições nos dados.

- O estudo do desenho também afeta a escolha dos modelos de probabilidade utilizados para descrever os dados.

- Veremos três exemplos a seguir.

Exemplo 1: Estudo transversal de melanoma maligno

- Esses dados são provenientes de um estudo transversal com pacientes diagnosticados com uma forma de câncer de pele chamada melanoma maligno.
- Para uma amostra de $n = 400$ pacientes, foram registrados o local do tumor e o seu tipo.
- Os dados, números de pacientes em cada combinação de tipo e local do tumor, estão apresentados na tabela a seguir.
- A questão de interesse é verificar se existe associação entre o tipo e o local do tumor.

Exemplo 1: Estudo transversal de melanoma maligno

Tipo de tumor	**Local**			
	Cabeça e pescoço	**Tronco**	**Extremidades**	**Total**
Lentigo melanótico de Hutchinson	22	2	10	34
Melanoma de disseminação superficial	16	54	115	185
Nodular	19	33	73	125
Indeterminado	11	17	28	56
Total	68	106	226	400

- A tabela a seguir mostra os dados exibidos como percentuais dos totais de linha e coluna.

Exemplo 1: Estudo transversal de melanoma maligno

Tipo de tumor	Local			
	Cabeça e pescoço	Tronco	Extremidades	Total
Porcentagens por linha				
Lentigo melanótico de Hutchinson	64,7	5,9	29,4	100
Melanoma de disseminação superficial	8,6	29,2	62,2	100
Nodular	15,2	26,4	58,4	100
Indeterminado	19,6	30,4	50,0	100
Todos os tipos	17,0	26,5	56,5	100
Porcentagens por coluna				
Lentigo melanótico de Hutchinson	32,4	1,9	4,4	8,50
Melanoma de disseminação superficial	23,5	50,9	50,9	46,25
Nodular	27,9	31,1	32,3	31,25
Indeterminado	16,2	16,0	12,4	14,00
Todos os tipos	100,0	99,9	100,0	100,0

Exemplo 1: Estudo transversal de melanoma maligno

- Parece que o *lentigo melanótico de Hutchinson* é mais comum na cabeça e no pescoço, mas há pouca evidência de associação entre os demais tipos de tumor e o local.

- Seja Y_{jk} a frequência na célula (j, k), com $j = 1, \ldots, J$ e $k = 1, \ldots, K$.

- Neste exemplo, temos $J = 4$ linhas, $K = 3$ colunas e a restrição

$$\sum_{j=1}^{J} \sum_{k=1}^{K} Y_{jk} = n,$$

onde $n = 400$ é fixado pelo desenho do estudo.

Exemplo 1: Estudo transversal de melanoma maligno

- Se Y_{jk} são variáveis aleatórias independentes com distribuição Poisson, tais que

$$E(Y_{jk}) = \mu_{jk},$$

então a soma possui distribuição Poisson com parâmetro

$$E(n) = \mu = \sum_{j=1}^{J} \sum_{k=1}^{K} \mu_{jk}.$$

- A distribuição conjunta de probabilidade dos Y_{jk}, condicional à soma total n, é multinomial:

$$f(\mathbf{y} \mid n) = n! \prod_{j=1}^{J} \prod_{k=1}^{K} \frac{\theta_{jk}^{y_{jk}}}{y_{jk}!},$$

onde

$$\theta_{jk} = \frac{\mu_{jk}}{\mu}.$$

Exemplo 1: Estudo transversal de melanoma maligno

- A soma dos θ_{jk} é igual a 1, pois

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{jk} = \mu,$$

  e $0 < \theta_{jk} < 1$.

- Assim, θ_{jk} pode ser interpretado como a probabilidade de uma observação pertencer à célula (j, k) da tabela.

- O valor esperado de Y_{jk} é

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

- Utilizando a função de ligação logarítmica:

$$\log(\mu_{jk}) = \log(n) + \log(\theta_{jk}).$$

Exemplo 2: Ensaio controlado da vacina contra a gripe

- Em um estudo prospectivo de uma nova vacina contra a gripe, os pacientes foram alocados aleatoriamente em dois grupos: um recebeu a nova vacina e o outro recebeu um placebo salino.

- A variável resposta corresponde aos níveis de anticorpos inibidores de hemaglutinina medidos no sangue seis semanas após a vacinação.

- Os níveis de anticorpos foram categorizados em classes (por exemplo, "pequeno", "médio" e "grande").

- As frequências em cada linha da tabela a seguir são fixadas pelo desenho do estudo, pois correspondem ao número total de indivíduos em cada grupo de tratamento (35 e 38, respectivamente).

Exemplo 2: Ensaio controlado da vacina contra a gripe

Grupo	Resposta			
	Pequeno	**Médio**	**Grande**	**Total**
Placebo	25	8	5	38
Vacina	6	18	11	35

- Queremos saber se o padrão de respostas é o mesmo para cada grupo de tratamento.

Exemplo 2: Ensaio controlado da vacina contra a gripe

- Neste exemplo, os totais das linhas são fixados pelo desenho do estudo.
- Assim, a distribuição de probabilidade conjunta para cada linha é multinomial:

$$f(y_{j1}, y_{j2}, \ldots, y_{jK} \mid y_{j\cdot}) = y_{j\cdot}! \prod_{k=1}^{K} \frac{\theta_{jk}^{y_{jk}}}{y_{jk}!},$$

onde

$$y_{j\cdot} = \sum_{k=1}^{K} y_{jk} \quad \text{e} \quad \sum_{k=1}^{K} \theta_{jk} = 1.$$

- Portanto, a distribuição conjunta para todas as células da tabela é o produto das distribuições multinomiais correspondentes a cada linha.

Exemplo 2: Ensaio controlado da vacina contra a gripe

- A distribuição conjunta condicional aos totais das linhas é dada por:

$$f(\mathbf{y} \mid y_{1\cdot}, y_{2\cdot}, \ldots, y_{J\cdot}) = \prod_{j=1}^{J} y_{j\cdot}! \prod_{k=1}^{K} \frac{\theta_{jk}^{y_{jk}}}{y_{jk}!},$$

onde

$$\sum_{k=1}^{K} \theta_{jk} = 1 \quad \text{para cada linha } j.$$

- Neste caso,

$$E(Y_{jk}) = y_{j\cdot}\theta_{jk},$$

e,

$$\log E(Y_{jk}) = \log(\mu_{jk}) = \log(y_{j\cdot}) + \log(\theta_{jk}).$$

- Se o padrão de resposta for o mesmo para ambos os grupos, então

$$\theta_{jk} = \theta_{\cdot k}, \quad k = 1, \ldots, K.$$

Exemplo 3: Estudo de Úlceras e Uso de Aspirina

- Neste estudo retrospectivo do tipo caso-controle, um grupo de pacientes com úlcera foi comparado a um grupo controle composto por pacientes que não sabiam se tinham úlcera, mas que eram semelhantes aos casos quanto à idade, ao sexo e ao status socioeconômico.
- Os pacientes com úlcera foram classificados de acordo com o local da lesão: gástrica ou duodenal.
- O uso de aspirina foi verificado para todos os participantes do estudo.
- Os resultados estão apresentados na tabela a seguir.

Exemplo 3: Estudo de Úlceras e Uso de Aspirina

		Uso de aspirina		
		Não usuário	**Usuário**	**Total**
Úlcera gástrica				
	Controle	62	6	68
	Casos	39	25	64
Úlcera duodenal				
	Controle	53	8	61
	Casos	49	8	57

Exemplo 3: Estudo de Úlceras e Uso de Aspirina

- Trata-se de uma tabela de contingência $2 \times 2 \times 2$.
- Algumas questões de interesse:
  1. A úlcera gástrica está associada ao uso de aspirina?
  2. A úlcera duodenal está associada ao uso de aspirina?
  3. Existe associação entre o uso de aspirina e ambos os tipos de úlcera?
- Quando os dados são apresentados como porcentagens por linha (próxima tabela), observa-se que o uso de aspirina é mais comum entre pacientes com úlcera gástrica do que entre os controles.
- No entanto, esse padrão não parece ocorrer para úlcera duodenal.

Exemplo 3: Estudo de Úlceras e Uso de Aspirina

		Uso de aspirina		
		Não usuário	**Usuário**	**Total**
Úlcera gástrica				
	Controle	91	9	100
	Casos	61	39	100
Úlcera duodenal				
	Controle	87	13	100
	Casos	86	14	100

Exemplo 3: Estudo de Úlceras e Uso de Aspirina

- Seja $j = 1, 2$ para controle e casos, respectivamente; $k = 1, 2$ para úlcera gástrica e duodenal, respectivamente; e $l = 1, 2$ para não usuário e usuário de aspirina.

- Seja Y_{jkl} a frequência observada na categoria (j, k, l), com $j = 1, \ldots, J$, $k = 1, \ldots, K$ e $l = 1, \ldots, L$.

- Se o total marginal $y_{jk\cdot}$ é fixado, então a distribuição conjunta de Y_{jkl} é

$$f(\mathbf{y} \mid y_{11\cdot}, \ldots, y_{JK\cdot}) = \prod_{j=1}^{J} \prod_{k=1}^{K} y_{jk\cdot}! \prod_{l=1}^{L} \frac{\theta_{jkl}^{y_{jkl}}}{y_{jkl}!},$$

onde $\mathbf{y}$ é o vetor dos Y_{jkl} e

$$\sum_{l=1}^{L} \theta_{jkl} = 1, \quad \text{para } j = 1, \ldots, J \text{ e } k = 1, \ldots, K.$$

Exemplo 3: Estudo de Úlceras e Uso de Aspirina

- Esta é outra forma do produto de distribuições multinomiais.
- Neste caso,

$$E(Y_{jkl}) = \mu_{jkl} = y_{jk.}\theta_{jkl},$$

e, utilizando a função de ligação logarítmica,

$$\log \mu_{jkl} = \log y_{jk.} + \log \theta_{jkl}.$$

Modelos de Probabilidade para Tabelas de Contingência

- Os exemplos anteriores ilustram os principais modelos de probabilidade utilizados para tabelas de contingência.
- Seja $\mathbf{y}$ o vetor das frequências Y_i nas N células da tabela.
- Consideraremos o caso geral envolvendo:
  - modelo de Poisson,
  - modelo Multinomial,
  - produto de modelos multinomiais.

Modelo Poisson

- Se não há restrições sobre os Y_i, eles podem ser modelados como variáveis aleatórias independentes com

$$E(Y_i) = \mu_i.$$

- A distribuição conjunta de probabilidade é dada por

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

onde $\boldsymbol{\mu}$ é o vetor dos parâmetros μ_i.

Modelo Multinomial

- Se a única restrição é que a soma dos Y_i seja n, então pode-se utilizar a distribuição multinomial:

$$f(\mathbf{y}; \boldsymbol{\theta}, n) = n! \prod_{i=1}^{N} \frac{\theta_i^{y_i}}{y_i!},$$

onde

$$\sum_{i=1}^{N} \theta_i = 1, \qquad \sum_{i=1}^{N} y_i = n, \qquad E(Y_i) = n\theta_i.$$

- Para uma tabela de contingência bidimensional (como no exemplo do melanoma), com j representando as linhas e k as colunas, a hipótese mais comum é a de independência entre as variáveis de linha e coluna.

Modelo Multinomial

- A hipótese de independência entre linha (j) e coluna (k) implica que

$$\theta_{jk} = \theta_{j.}\,\theta_{.k},$$

onde $\theta_{j.}$ e $\theta_{.k}$ são as probabilidades marginais, com

$$\sum_j \theta_{j.} = 1 \quad \text{e} \quad \sum_k \theta_{.k} = 1.$$

- Essa hipótese pode ser testada comparando o ajuste de dois modelos lineares para

$$\mu_{jk} = E(Y_{jk}).$$

- Modelo geral:

$$\log \mu_{jk} = \log n + \log(\theta_{jk}).$$

- Modelo sob independência:

$$\log \mu_{jk} = \log n + \log \theta_{j.} + \log \theta_{.k}.$$

Modelo Multinomial Produto

- Se houver mais totais marginais fixos além do total global n, então produtos apropriados de distribuições multinomiais podem ser utilizados para modelar os dados.

- Por exemplo, para uma tabela tridimensional com J linhas, K colunas e L níveis, suponha que os totais das linhas sejam fixados em cada nível l.

- A distribuição conjunta dos Y_{jkl} é dada por

$$f(\mathbf{y} \mid y_{j\cdot l}) = \prod_{j=1}^{J} \prod_{l=1}^{L} y_{j\cdot l}! \prod_{k=1}^{K} \frac{\theta_{jkl}^{y_{jkl}}}{y_{jkl}!},$$

onde

$$\sum_{k=1}^{K} \theta_{jkl} = 1 \quad \text{para cada combinação } (j, l).$$

Modelo Multinomial Produto

- Neste caso,

$$E(Y_{jkl}) = y_{j\cdot l}\,\theta_{jkl}.$$

- Se apenas os totais por nível l forem fixados, isto é, $y_{\cdot\cdot l}$ fixos, então a distribuição conjunta é

$$f(\mathbf{y} \mid y_{\cdot\cdot l}) = \prod_{l=1}^{L} y_{\cdot\cdot l}! \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\theta_{jkl}^{y_{jkl}}}{y_{jkl}!},$$

onde

$$\sum_{j=1}^{J}\sum_{k=1}^{K} \theta_{jkl} = 1, \quad \text{para } l = 1, \ldots, L.$$

- Nesse caso,

$$E(Y_{jkl}) = y_{\cdot\cdot l}\,\theta_{jkl}.$$

Modelo Log-Linear

- Todos os modelos de probabilidade descritos anteriormente são baseados na distribuição de Poisson e, em todos os casos, $E(Y_i)$ pode ser escrita como um produto de parâmetros e outros termos.

- A função de ligação canônica da distribuição Poisson, a função logarítmica, produz um componente linear:

$$\log E(Y_i) = \text{constante} + \mathbf{x}_i^\top \boldsymbol{\beta}.$$

- O termo *modelo log-linear* é utilizado para descrever essa classe de modelos lineares generalizados aplicados a tabelas de contingência.

Modelo Log-Linear

- No Exemplo 1 (melanoma), se não houver associação entre local e tipo de tumor, isto é, se as variáveis forem independentes, então a probabilidade conjunta pode ser escrita como o produto das probabilidades marginais:

$$\theta_{jk} = \theta_{j\cdot}\,\theta_{\cdot k}, \quad j = 1, \ldots, J,\ k = 1, \ldots, K.$$

- A hipótese de independência pode ser testada comparando o modelo aditivo (na escala logarítmica)

$$\log E(Y_{jk}) = \log n + \log \theta_{j\cdot} + \log \theta_{\cdot k}, \tag{1}$$

com o modelo geral

$$\log E(Y_{jk}) = \log n + \log \theta_{jk}. \tag{2}$$

Modelo Log-Linear

- Isso é análogo à análise de variância para um experimento com dois fatores sem replicação.

- A equação (2) pode ser escrita como o modelo saturado:

$$\log E(Y_{jk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}.$$

- Já a equação (1) corresponde ao modelo de independência:

$$\log E(Y_{jk}) = \mu + \alpha_j + \beta_k.$$

- Como o termo $\log n$ está presente em todos os modelos, o modelo mínimo (modelo nulo) é:

$$\log E(Y_{jk}) = \mu.$$

Modelo Log-Linear

- No Exemplo 2 (vacina contra a gripe),

$$E(Y_{jk}) = y_{j.}\theta_{jk}$$

se as distribuições de resposta descritas pelos θ_{jk} diferem entre os grupos j, ou

$$E(Y_{jk}) = y_{j.}\theta_{.k}$$

se a distribuição é a mesma para todos os grupos.

- A hipótese de homogeneidade das distribuições pode ser testada comparando: Modelo geral (com interação):

$$\log E(Y_{jk}) = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk},$$

correspondente a $E(Y_{jk}) = y_{j.}\theta_{jk}$, e o modelo de homogeneidade

$$\log E(Y_{jk}) = \mu + \alpha_j + \beta_k,$$

correspondente a $E(Y_{jk}) = y_{j.}\theta_{.k}$.

Modelo Log-Linear

- O modelo mínimo para esses dados é

$$\log E(Y_{jk}) = \mu + \alpha_j,$$

pois os totais das linhas (índice j) são fixados pelo desenho do estudo.

- A especificação dos componentes lineares em modelos log-lineares apresenta muitas semelhanças com a formulação de modelos ANOVA.

- Os modelos são hierárquicos: se um termo de ordem superior (por exemplo, uma interação) é incluído no modelo, então todos os termos relacionados de ordem inferior também devem ser incluídos.

Modelo Log-Linear

- Se a interação de primeira ordem $(\alpha\beta)_{jk}$ é incluída no modelo, então os efeitos principais α_j, β_k e a constante μ também devem ser incluídos.
- De forma semelhante, se uma interação de segunda ordem $(\alpha\beta\gamma)_{jkl}$ é incluída, então também devem ser incluídas as interações de primeira ordem

$$(\alpha\beta)_{jk}, \quad (\alpha\gamma)_{jl}, \quad (\beta\gamma)_{kl}.$$

- Como os modelos log-lineares são especificados de maneira análoga aos modelos ANOVA, eles envolvem muitos parâmetros. Assim, restrições de identificabilidade, como as restrições de soma zero ou parametrização por corner point, são necessárias.
- Em tabelas de contingência, as principais questões estão quase sempre relacionadas à associação entre variáveis.
- Assim, em modelos log-lineares, os termos de interesse primário são as interações que envolvem duas ou mais variáveis.

Inferência em Modelo Log-Linear

- Embora três tipos de distribuições de probabilidade possam ser utilizados para descrever dados de tabelas de contingência (Poisson, Multinomial e produto de Multinomiais), pode-se demonstrar que, para qualquer modelo log-linear, os estimadores de máxima verossimilhança são os mesmos sob todas essas distribuições.

- Isso implica que, para fins de estimação, a distribuição de Poisson pode ser sempre utilizada.

- Como a distribuição de Poisson pertence à família exponencial e as restrições paramétricas podem ser incorporadas ao componente linear, todos os métodos usuais para Modelos Lineares Generalizados (MLGs) podem ser aplicados.

Modelo Log-Linear: Avaliação e Testes

- A adequação do modelo pode ser avaliada por meio das estatísticas de bondade de ajuste, como a deviance D e o qui-quadrado de Pearson X^2 (e, em alguns casos, estatística C ou pseudo R^2).

- Informações adicionais sobre o ajuste podem ser obtidas a partir da análise dos resíduos de Pearson ou dos resíduos de deviance.

- Testes de hipóteses podem ser realizados comparando as estatísticas de bondade de ajuste de um modelo mais geral (hipótese alternativa) com as de um modelo mais simples encaixado (hipótese nula).