Uma introdução à análise de dados categorizados com respostas completas ou com omissão utilizando a biblioteca de rotinas `Catdata` para o `R`

Frederico Zanqueta Poleto (fred@poleto.com), IME–USP

com colaboração de

Julio da Motta Singer (jmsinger@ime.usp.br), IME–USP

Carlos Daniel Paulino (dpaulino@math.ist.utl.pt), IST–UTL

Este trabalho recebeu apoio financeiro do CNPq e FAPESP.

Minicurso para o LEG–UFPR, 23/03/2007

Referências

PAULINO, C.D.M. e SINGER, J.M. (2006). *Análise de dados categorizados*. Edgard Blücher.

SINGER, J.M. (2003). *Análise de dados categorizados*. Texto do minicurso apresentado no XIII Simposio de Estadistica "Estadística en Ciencias de la Salud". Disponível em `http://www.ime.usp.br/~jmsinger/Textos/ADC-Colombia2.pdf`

AGRESTI, A. (2002). *Categorical data analysis*. 2ª ed. John Wiley & Sons.

FLEISS, J.L., LEVIN, B. e PAIK, M.C. (2003). *Statistical methods for rates and proportions*. 3ª ed. John Wiley & Sons.

BISHOP, Y.M.M., FIENBERG, S.E. e HOLLAND, P.W. (1975). *Discrete multivariate analysis: theory and practice*. The MIT Press.

Referências

Disponível em `http://www.poleto.com/missing.html`:

- POLETO, F.Z. (2006). *Análise de dados categorizados com omissão*. Dissertação de mestrado. IME–USP.
- ——— **(2007).** ***Comandos (em R) para reproduzir as análises de exemplos do livro Análise de Dados Categorizados de Paulino e Singer (2006)*****. Manuscrito não-publicado.**
- POLETO, F.Z., SINGER, J.M. e PAULINO, C.D. (2007a). *A product-multinomial framework for categorical data analysis with missing responses*. Relatório técnico RT-MAE-2007-07. IME–USP.
- ——— (2007b). Comparing diagnostic tests with missing data. Submetido para publicação. (Código do R para reproduzir as análises do manuscrito disponíveis)
- ——— (2007c). *Analyzing categorical data with complete or missing responses using the Catdata package*. Vinheta para o R não-publicada.
- THOMPSON, L.A. (2007). *S-Plus (and R) manual to accompany Agresti's "Categorical Data Analysis"(2002) 2^{nd} edition*. Disponível em `https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf`

Tabelas de contingência $S \times R$

Subpopulação	Categorias de resposta						Total
	1	2	$\cdots$	r	$\cdots$	R	
1	n_{11}	n_{12}	$\cdots$	n_{1r}	$\cdots$	n_{1R}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	n_{2r}	$\cdots$	n_{2R}	n_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
s	n_{s1}	n_{s2}	$\cdots$	n_{sr}	$\cdots$	n_{sR}	n_{s+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
S	n_{S1}	n_{S2}	$\cdots$	n_{Sr}	$\cdots$	n_{SR}	n_{S+}
Total							n_{++}

questões de interesse
planejamento amostral $\longrightarrow$ variáveis explicativas $\times$ respostas

Problema da intenção de voto

- Sondagens realizadas sobre as intenções de voto de 445 pessoas em duas entrevistas espaçadas de um mês
- Objetivo: avaliar se as mudanças na intenção de voto são iguais nos dois sentidos

1ª sondagem	2ª sondagem		
	partido A	partido B	indeciso
partido A	192	1	5
partido B	2	146	5
indeciso	11	12	71

$$S = 1 \quad \text{e} \quad R = 3 \times 3 = 9$$

1ª	partido A			partido B			indeciso		
2ª	A	B	ind.	A	B	ind.	A	B	ind.
	192	1	5	2	146	5	11	12	71

Paulino e Singer (2006), Exemplo 3.1 / 8.1 / 11.2

Estudo de suscetibilidade à cárie dentária

- 97 crianças de 11 a 13 anos de uma escola pública
- Objetivo: avaliar se o método simplificado (mais barato) é tão eficaz quanto o método convencional (de custos elevados)

Método	Método convencional		
simplificado	baixo	médio	alto
baixo	11	5	0
médio	14	34	7
alto	2	13	11

$$S = 1 \quad \text{e} \quad R = 3 \times 3 = 9$$

Simpl.	baixo			médio			alto		
Conv.	baixo	médio	alto	baixo	médio	alto	baixo	médio	alto
	11	5	0	14	34	7	2	13	11

Paulino e Singer (2006), Exemplo 1.2 / 3.2 / 8.2 / 10.2 / 11.3 / 11.12

Problema do tamanho da ninhada

- Estudo de fertilidade de ovelhas de vários rebanhos
- Objetivo: avaliar a influência da fazenda e da raça no tamanho da ninhada

Fazenda	Raça	Tamanho da ninhada				Total
		0	1	2	≥ 3	
1	A	10	21	96	23	150
	B	4	6	28	8	46
	C	9	7	58	7	81
2	A	8	19	44	1	72
	B	5	17	56	1	79
	C	1	5	20	2	28
3	A	22	95	103	4	224
	B	18	49	62	0	129
	C	4	12	16	2	34

$$S = 3 \times 3 = 9 \quad \text{e} \quad R = 4$$

Paulino e Singer (2006), Exemplo 1.9 / 3.3 / 8.3 / 11.4

Problema da anemia

- Estudo da FSP–USP com 128 crianças com 4 meses de idade em região com raras situações de desnutrição e miséria extrema
- Objetivo: avaliar se o aleitamento materno e a anemia estão associadas

Anemia	Aleitamento	
	apenas materno	misto
sim	3	25
não	32	68

$$S = 1 \quad \text{e} \quad R = 2 \times 2 = 4$$

Anemia	sim		não	
Aleitamento	materno	misto	materno	misto
	3	25	32	68

Paulino e Singer (2006), Exemplo 9.1 / 11.5

Estudo da satisfação com o emprego

- 96 homens dos EUA foram sondados
- Objetivo: avaliar se o salário influencia a satisfação com o emprego

Renda anual (US$)	Satisfação com o emprego			
	muito insatisfeito	um pouco insatisfeito	um pouco satisfeito	muito satisfeito
$<$15,000	1	3	10	6
15,000$-$25,000	2	3	10	7
25,000$-$40,000	1	6	14	12
$>$40,000	0	1	9	11

$$S = 1 \quad \text{e} \quad R = 4 \times 4 = 16$$

Renda	$<$15,000				$\cdots$	$>$40,000			
Satisf.	MI	PI	PS	MS	$\cdots$	MI	PI	PS	MS
	1	3	10	6	$\cdots$	0	1	9	11

Agresti (2002), Seção 2.4.3

Problema da fobia em alcoólatras

- Estudo realizado com 93 alcoólatras
- Objetivo: avaliar se a presença de fobia, o consumo diário de álcool e a situação profissional estão relacionadas

Situação profissional	Uso diário de álcool	Fobia	
		sim	não
sem emprego	sim	10	24
	não	6	12
com emprego	sim	13	17
	não	4	7

$$S = 1 \quad \text{e} \quad R = 2 \times 2 \times 2 = 8$$

Emprego	não				sim			
Uso diário	sim		não		sim		não	
Fobia	S	N	S	N	S	N	S	N
	10	24	6	12	13	17	4	7

Paulino e Singer (2006), Exemplo 1.11 / 9.4

Problema de uso de fio dental

- 30 crianças de cada sexo e faixa etária (5-8 e 9-12 anos) foram selecionadas de uma escola da rede pública do município de SP
- Objetivo: avaliar se o sexo e a faixa etária influenciam a freqüência e a habilidade no uso do fio dental

Sexo	Faixa etária	Freqüência	Habilidade	
			inábil	razoável
masc.	5-8	insuficiente	19	5
		boa	4	2
	9-12	insuficiente	5	8
		boa	0	17
fem.	5-8	insuficiente	11	6
		boa	7	6
	9-12	insuficiente	2	5
		boa	1	22

$$S = 2 \times 2 = 4 \quad \text{e} \quad R = 2 \times 2 = 4$$

Paulino e Singer (2006), Exemplo 1.5 / 9.11

Problema de uso de fio dental

Sexo	Faixa etária	Freqüência	Habilidade	
			inábil	razoável
masc.	5-8	insuficiente	19	5
		boa	4	2
	9-12	insuficiente	5	8
		boa	0	17
fem.	5-8	insuficiente	11	6
		boa	7	6
	9-12	insuficiente	2	5
		boa	1	22

Sexo	Faixa etária	Freq. Habil.	insuficiente		boa	
			inábil	razoável	inábil	razoável
masc.	5-8		19	5	4	2
	9-12		5	8	0	17
fem.	5-8		11	6	7	6
	9-12		2	5	1	22

$$S = 2 \times 2 = 4 \quad \text{e} \quad R = 2 \times 2 = 4$$

Problema da complicação pulmonar

- 1162 pacientes tiveram os graus de complicação pulmonar pré e pós-operatório avaliados
- Objetivo: comparar os riscos de complicação pulmonar do período pós-operatório entre os níveis da avaliação do pré-operatório

Avaliação de complicação pulmonar

Pré-	Pós-operatório	
-operatório	sem complicação	com complicação
baixo	737	48
moderado	243	74
alto	39	21

$$S = 3 \quad \text{e} \quad R = 2$$

Paulino e Singer (2006), Exemplo 6.5 / 10.4 / 11.10

Modelos probabilísticos

A escolha do modelo probabilístico para os dados depende do planejamento e do objetivo do estudo

Consideraremos 3 estratégias de obtenção dos dados de uma *pesquisa de intenção de voto*, cujo interesse é avaliar a relação entre a *opinião de eleitores sobre um determinado candidato* (X_1) e sua *faixa etária* (X_2)

Estratégia I

Entrevistar tantas pessoas quanto possível, por exemplo, em 4 horas

Faixa etária	Opinião		Total
	favorável	desfavorável	
< 40	43	41	
≥ 40	25	70	
Total			179

Estratégia I - suposições

Suposições sobre o nº de transeuntes com < 40 anos favoráveis ao candidato que passa no sítio em que se vai colher a amostra:

- num determinado intervalo de tempo, o nº desses transeuntes é independente do nº de transeuntes com as mesmas características que passa em qualquer outro intervalo de tempo disjunto daquele;
- a distribuição daquele nº de transeuntes só depende do comprimento do intervalo de tempo considerado e não do seu instante inicial;
- a probabilidade de passagem de um daqueles transeuntes num intervalo de tempo suficientemente pequeno (*e.g.*, um segundo) é aproximadamente proporcional ao comprimento do intervalo, com constante de proporcionalidade λ_{11};
- a probabilidade de que dois ou mais daqueles transeuntes passem simultaneamente num intervalo de tempo suficientemente pequeno é desprezável.

Estratégia I - modelo probabilístico

Faixa etária	Opinião		Total
	favorável ($j=1$)	desfavorável ($j=2$)	
<40 ($i=1$)	43	41	
≥ 40 ($i=2$)	25	70	
Total			179

$$n_{ij} \sim \text{Poisson}(\mu_{ij}), \quad i,j=1,2, \quad \text{indep.},$$

em que $\mu_{ij} = m\lambda_{ij}, \quad m = 4 \times 3600s = 14400s$

$\therefore \ \mathbf{N} = (n_{11}, n_{12}, n_{21}, n_{22})' \sim$ Produto de distrib.de Poisson($\boldsymbol{\mu}$) com f.p.

$$f(\mathbf{N}|\boldsymbol{\mu}) = \prod_{i=1}^{2}\prod_{j=1}^{2} \frac{\mathrm{e}^{-\mu_{ij}}\,\mu_{ij}^{n_{ij}}}{n_{ij}!}, \quad n_{ij} \in \mathbb{N}_o, \ i,j=1,2$$

em que $\boldsymbol{\mu} = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22})'$, com $\mu_{ij} \in \mathbb{R}^+, \ i,j=1,2$

[note que $n_{++} \sim$ Poisson(μ_{++}), $n_{i+} \sim$ Poisson(μ_{i+}), $n_{+j} \sim$ Poisson(μ_{+j})]

Estratégia I - hipótese de interesse

Faixa etária	Opinião		Total
	favorável ($j=1$)	desfavorável ($j=2$)	
<40 ($i=1$)	μ_{11}	μ_{12}	μ_{1+}
≥ 40 ($i=2$)	μ_{21}	μ_{22}	μ_{2+}
Total	μ_{+1}	μ_{+2}	μ_{++}

Hipótese de interesse: a proporção de apoiantes entre os indivíduos mais jovens é a mesma que existe entre as pessoas menos jovens

$$H_I : \frac{\mu_{11}}{\mu_{1+}} = \frac{\mu_{21}}{\mu_{2+}} \left(= \frac{\mu_{+1}}{\mu_{++}} \right)$$

Pode-se reescrever essa hipótese como

$$H_I : \mu_{ij} = \frac{\mu_{i+} \times \mu_{+j}}{\mu_{++}}, \quad i,j = 1,2$$

Estratégia II / suposições

Fixar antecipadamente o número n_{++} de pessoas a entrevistar e selecioná-las de um modo aleatório. Por ex., $n_{++} = 200$.

Faixa etária (X_1)	Opinião (X_2)		Total
	favorável ($j = 1$)	desfavorável ($j = 2$)	
$<$ 40 ($i = 1$)	50	48	
$\geq$ 40 ($i = 2$)	26	76	
Total			200

θ_{ij}: probabilidade de um indivíduo apresentar a característica (i, j), considerada constante para todo o indivíduo da população em estudo, *i.e.*, $\theta_{ij} = P\left(X_{1k} = i, X_{2k} = j\right)$, $k = 1, \ldots, n_{++}$

Seja $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})'$, então $\mathbf{1}_4'\boldsymbol{\theta} = \sum_{i,j} \theta_{ij} = 1$

Estratégia II - suposições / modelo prob.

Associe-se ao indivíduo k o vetor $\mathbf{W}_k = (W_{k11}, W_{k12}, W_{k21}, W_{k22})'$, em que $W_{kij} = 1$ e $W_{ki'j'} = 0$ para $i' \neq i$ e/ou $j' \neq j$ quando $X_{1k} = i$ e $X_{2k} = j$

Logo, $\mathbf{W}_k \in \{(1,0,0,0),\ (0,1,0,0),\ (0,0,1,0),\ (0,0,0,1)\}$

e $\mathbf{W}_k \sim$ Bernoulli multivariada($\boldsymbol{\theta}$), $k = 1, \ldots, n_{++}$ (identicamente distrib.)

Assumindo que esses vetores $\mathbf{W}_k$, $k = 1, \ldots, n_{++}$ são independentes,

segue-se $\mathbf{N} = \sum_{k=1}^{n_{++}} \mathbf{W}_k \sim$ Multinomial($n_{++}, \boldsymbol{\theta}$) com f.p.

$$f\left(\mathbf{N}|n_{++}, \boldsymbol{\theta}\right) = n_{++}! \prod_{i=1}^{2} \prod_{j=1}^{2} \frac{\theta_{ij}^{n_{ij}}}{n_{ij}!},$$

com $\mathbf{1}_4'\mathbf{N} = n_{++}$, $\mathbf{1}_4'\boldsymbol{\theta} = 1$

[note que $n_{i+} \sim$ Binomial(n_{++}, θ_{i+}), $n_{+j} \sim$ Binomial(n_{++}, θ_{+j})]

Estratégia II - hipótese de interesse

Faixa etária (X_1)	Opinião (X_2)		Total
	favorável ($j=1$)	desfavorável ($j=2$)	
<40 ($i=1$)	θ_{11}	θ_{12}	θ_{1+}
≥ 40 ($i=2$)	θ_{21}	θ_{22}	θ_{2+}
Total	θ_{+1}	θ_{+2}	1

Hipótese de interesse: independência estocástica entre X_1 e X_2

$$H_{II} : \theta_{ij} = \theta_{i+} \times \theta_{+j}, \quad i,j = 1,2$$

Estratégia III / suposições

Fixar antecipadamente o número n_{i+} de indivíduos de cada faixa etária.

Por ex., $n_{1+} = n_{2+} = 100$.

Faixa etária (X_1)	Opinião (X_2)		Total
	favorável ($j = 1$)	desfavorável ($j = 2$)	
< 40 ($i = 1$)	54	46	100
≥ 40 ($i = 2$)	30	70	100
Total			200

$\theta_{j(i)}$: probabilidade de um indivíduo que possui a categoria i de X_1 ter a classificação j em X_2, *i.e.*,

$\theta_{j(i)} = P\left(X_{2k} = j | X_{1k} = i\right)$, $k = 1, \ldots, n_{++}$, $i = 1, 2$

Estratégia III - modelo probabilístico

Seja $\boldsymbol{\theta} = \left(\theta_{1(1)}, \theta_{2(1)}, \theta_{1(2)}, \theta_{2(2)}\right)'$

Argumentos similares levam a

$\mathbf{N} \sim$ Produto de distribuições Multinomiais($n_{i+}, \theta_{j(i)}$), $i = 1, 2$, com f.p.

$$f\left(\mathbf{N}|\mathbf{N}_+, \boldsymbol{\theta}\right) = \prod_{i=1}^{2} \left[n_{i+}! \prod_{j=1}^{2} \frac{\theta_{j(i)}^{n_{ij}}}{n_{ij}!} \right],$$

em que $\mathbf{N}_+ = (n_{1+}, n_{2+})'$, $\sum_j \theta_{j(i)} = 1,\ i = 1, 2$

Estratégia III - hipótese de interesse

Faixa etária (X_1)	Opinião (X_2)		Total
	favorável ($j=1$)	desfavorável ($j=2$)	
<40 ($i=1$)	$\theta_{1(1)}$	$\theta_{2(1)}$	1
≥ 40 ($i=2$)	$\theta_{1(2)}$	$\theta_{2(2)}$	1

Hipótese de interesse: homogeneidade das distribuições Multinomiais

$$H_{III} : \theta_{1(1)} = \theta_{1(2)}$$

Relações entre modelos

Pode-se mostrar (veja Paulino e Singer, 2006, p.27; Singer, 2003, p.14)

- O **modelo Multinomial** com parâmetros $\theta_{ij} = \mu_{ij}/\mu_{++}$ **pode ser obtido** a partir do **modelo Produto de distribuições de Poisson** por condicionamento no total da tabela n_{++}
- O **modelo produto de distribuições Multinomiais** com parâmetros $\theta_{j(i)} = \mu_{ij}/\mu_{i+}$ **pode ser obtido** a partir do **modelo Produto de distribuições de Poisson** por condicionamento nos totais marginais n_{i+}
- O **modelo produto de distribuições Multinomiais** com parâmetros $\theta_{j(i)} = \theta_{ij}/\theta_{i+}$ pode ser obtido a partir do **modelo Multinomial** por condicionamento nos totais marginais n_{i+}

Esses resultados permitem que a classificação de algumas variáveis como explicativas ou fatores seja feita a posteriori, por condicionamento.

Medidas de associação

Fator de risco	Estado do paciente		Total
	sem doença	doente	
não exposto	$1-\pi_0$	π_0	1
exposto	$1-\pi_1$	π_1	1

- π_0 (π_1): proporção de pacientes **não expostos** (**expostos**) ao fator de risco que apresentaram a doença
- Diferença de proporções: $d = \pi_1 - \pi_0 \rightarrow \pi_1 = d + \pi_0$
  há um aumento de d na proporção de doentes atribuível à exposição ao fator de risco
- Risco relativo: $r = \pi_1/\pi_0 \rightarrow \pi_1 = r\pi_0 \rightarrow \pi_1 = \pi_0 + (r-1)\pi_0$
  a proporção de doentes entre os expostos ao fator de risco é r vezes (ou, $r-1$ maior) a proporção de doentes entre os não expostos ao fator de risco
- $\pi_0/(1-\pi_0)$: **chance** de um indivíduo ser doente *vs.* não doente quando **não exposto** ao fator de risco
  $\pi_1/(1-\pi_1)$: **chance** de um indivíduo ser doente *vs.* não doente quando **exposto** ao fator de risco

  Razão de chances: $\omega = \dfrac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \rightarrow \dfrac{\pi_1}{1-\pi_1} = \omega\dfrac{\pi_0}{1-\pi_0}$

Medidas de associação - exemplos

Fator de risco	Estado do paciente		Total
	sem doença	doente	
não exposto	$1-\pi_0$	π_0	1
exposto	$1-\pi_1$	π_1	1

- Diferença de proporções: $d = \pi_1 - \pi_0$
- Risco relativo: $r = \pi_1/\pi_0$
- Razão de chances: $\omega = \dfrac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$
- Exemplo 1: $\pi_0 = 0.42, \pi_1 = 0.44 \quad \rightarrow \quad d = 0.02, r \cong 1.05, \omega = 1.09$
- Exemplo 2: $\pi_0 = 0.02, \pi_1 = 0.04 \quad \rightarrow \quad d = 0.02, r = 2.00, \omega \cong 2.04$
- $\omega = \dfrac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = r\dfrac{1-\pi_0}{1-\pi_1} \longrightarrow r$, quando π_0 e $\pi_1 \longrightarrow 0$

Em geral, se trabalha com linearizações de r e ω (modelos log-lineares)

- $\log r = \log \pi_1 - \log \pi_0$
- $\log \omega = \log \pi_1 - \log \pi_0 - \log(1-\pi_1) + \log(1-\pi_0)$

Estudos prospectivo *vs.* retrospectivo

Estudo prospectivo

Fator de risco	Estado do paciente		Total
	sem doença	doente	
não exposto	$1-\pi_0$	π_0	1
exposto	$1-\pi_1$	π_1	1

Estudo retrospectivo / caso-controle

Fator de risco	Estado do paciente	
	sem doença	doente
não exposto	$1-p_0$	$1-p_1$
exposto	p_0	p_1
Total	1	1

Utilizando o Teorema de Bayes pode-se demonstrar que

$$\omega = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

Modelo produto de Multinomiais

Subpopulação	Categorias de resposta						Total
	1	2	$\cdots$	r	$\cdots$	R	
1	$\theta_{1(1)}$	$\theta_{2(1)}$	$\cdots$	$\theta_{r(1)}$	$\cdots$	$\theta_{R(1)}$	1
2	$\theta_{1(2)}$	$\theta_{2(2)}$	$\cdots$	$\theta_{r(2)}$	$\cdots$	$\theta_{R(2)}$	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
s	$\theta_{1(s)}$	$\theta_{2(s)}$	$\cdots$	$\theta_{r(s)}$	$\cdots$	$\theta_{R(s)}$	1
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
S	$\theta_{1(S)}$	$\theta_{2(S)}$	$\cdots$	$\theta_{r(S)}$	$\cdots$	$\theta_{R(S)}$	1

$$\boldsymbol{\theta} = \left(\theta_{1(1)}, \theta_{2(1)}, \ldots, \theta_{R(S)}\right)' = \left(\boldsymbol{\theta}_s', s = 1, \ldots, S\right)'$$

$$\boldsymbol{\theta}_s = \left(\theta_{1(s)}, \theta_{2(s)}, \ldots, \theta_{R(s)}\right)', \ \mathbf{1}_R'\boldsymbol{\theta}_s = 1, \ s = 1, \ldots, S,$$

EMV para o modelo saturado

Maximizando

$$L\left(\boldsymbol{\theta}|\mathbf{N}\right) = \prod_{s=1}^{S}\left[n_{s+}!\prod_{r=1}^{R}\frac{\theta_{r(s)}^{n_{sr}}}{n_{sr}!}\right]$$

com relação a $\boldsymbol{\theta}$, obtém-se os EMVs $\{\hat{\boldsymbol{\theta}}_s\}$ de $\{\boldsymbol{\theta}_s\}$ que coincidem com as proporções amostrais:

$$\hat{\theta}_{r(s)} = \frac{n_{sr}}{n_{s+}}, \quad r = 1,\ldots,R, \longrightarrow \hat{\boldsymbol{\theta}}_s = \frac{1}{n_{s+}}\mathbf{N}_s, \quad s = 1,\ldots,S,$$

$\mathbf{N}_s = \left(n_{1(s)}, n_{2(s)}, \ldots, n_{R(s)}\right)'$

Matriz de covariâncias

Recorde que

$$Var(n_{sr}) = n_{s+}\theta_{r(s)}(1-\theta_{r(s)})$$
$$Cov(n_{sr}, n_{sr'}) = -n_{s+}\theta_{r(s)}\theta_{r'(s)}, \quad r \neq r'$$
$$Cov(n_{sr}, n_{s'r'}) = 0, \quad s \neq s'$$

Portanto,

$$Var(\hat{\theta}_{sr}) = \frac{1}{n_{s+}}\theta_{r(s)}(1-\theta_{r(s)})$$
$$Cov(\hat{\theta}_{sr}, \hat{\theta}_{sr'}) = -\frac{1}{n_{s+}}\theta_{r(s)}\theta_{r'(s)}, \quad r \neq r'$$
$$Cov(\hat{\theta}_{sr}, \hat{\theta}_{s'r'}) = 0, \quad s \neq s'$$

Logo, $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ é uma matriz bloco diagonal com blocos iguais a

$$\mathbf{V}_{\hat{\boldsymbol{\theta}}_s} = \frac{1}{n_{s+}}\left(\mathbf{D}_{\boldsymbol{\theta}_s} - \boldsymbol{\theta}_s\boldsymbol{\theta}_s'\right)$$

Modelos estruturais

Em geral, queremos modelar $\boldsymbol{\theta}$ por meio de estruturas não saturadas com a finalidade de dar respostas a questões de interesse.

Por exemplo, por meio de

- modelos lineares

$$M_L : \mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$$

- modelos log-lineares

$$M_{LL} : \mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{X}_L\boldsymbol{\beta}$$

- modelos funcionais lineares

$$M_F : \mathbf{F}(\boldsymbol{\theta}) = \mathbf{X}\boldsymbol{\beta}$$

EMV para modelos não saturados

- Pode-se incorporar as estruturas lineares e log-lineares reescrevendo $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\beta})$
- Maximizando

$$L\left(\boldsymbol{\theta}(\boldsymbol{\beta})|\mathbf{N}\right) = \prod_{s=1}^{S}\left\{n_{s+}!\prod_{r=1}^{R}\frac{\left[\theta_{r(s)}(\boldsymbol{\beta})\right]^{n_{sr}}}{n_{sr}!}\right\}$$

  com relação a $\boldsymbol{\beta}$, obtém-se o EMV $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$

- Em alguns casos há soluções explícitas para o EMV de $\boldsymbol{\beta}$, mas em geral costuma-se recorrer a métodos iterativos
- $\hat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}$ pode ser obtida por meio da inversa da estimativa da matriz de informação de Fisher

EMV para modelos não saturados

- A partir de $\hat{\boldsymbol{\beta}}$, pelo princípio da invariância, obtêm-se os EMVs (sob o modelo)
  - de $\mathbf{A}\boldsymbol{\theta}$ ou de $\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta})$ com $\mathbf{X}\hat{\boldsymbol{\beta}}$ ou $\mathbf{X}_L\hat{\boldsymbol{\beta}}$
  - de $\boldsymbol{\theta}(\boldsymbol{\beta})$ com $\boldsymbol{\theta}(\hat{\boldsymbol{\beta}})$

- Por meio do método *delta* obtém-se as correspondentes estimativas das matrizes de covariância assintóticas

- Pode-se obter também as freqüências estimadas sob o modelo, *i.e.*,

$$\hat{\mathbf{N}}_s = E(\mathbf{N}_s|n_{s+}, \boldsymbol{\theta}_s(\hat{\boldsymbol{\beta}})) = n_{s+}\boldsymbol{\theta}_s(\hat{\boldsymbol{\beta}})$$

Testes de ajustamento dos modelos

Seja M (M_L ou M_{LL})

- Estatística de razão de verossimilhanças

$$Q_V(M) = -2 \ln \frac{L\left(\boldsymbol{\theta}(\hat{\boldsymbol{\beta}}) \,|\mathbf{N}\right)}{L\left(\hat{\boldsymbol{\theta}} \,|\mathbf{N}\right)} = -2\mathbf{N}' \left\{ \mathbf{ln}\left[\boldsymbol{\theta}(\hat{\boldsymbol{\beta}})\right] - \mathbf{ln}\left[\hat{\boldsymbol{\theta}}\right] \right\}$$

- Estatística de Pearson

$$Q_P(M) = \sum_{s=1}^{S} \sum_{r=1}^{R} \frac{\left(n_{sr} - n_{s+}\theta_{r(s)}(\hat{\boldsymbol{\beta}})\right)^2}{n_{s+}\theta_{r(s)}(\hat{\boldsymbol{\beta}})} = \left(\mathbf{N} - \hat{\mathbf{N}}\right)' \mathbf{D}_{\hat{\mathbf{N}}}^{-1} \left(\mathbf{N} - \hat{\mathbf{N}}\right)$$

- Estatística de Neyman

$$Q_N(M) = \sum_{s=1}^{S} \sum_{r=1}^{R} \frac{\left(n_{sr} - n_{s+}\theta_{r(s)}(\hat{\boldsymbol{\beta}})\right)^2}{n_{sr}} = \left(\mathbf{N} - \hat{\mathbf{N}}\right)' \mathbf{D}_{\mathbf{N}}^{-1} \left(\mathbf{N} - \hat{\mathbf{N}}\right)$$

Testes de ajustamento dos modelos

- Estatística de Wald

$$Q_W(M_L) = \left(\mathbf{U}\mathbf{A}\hat{\boldsymbol{\theta}}\right)' \left[\mathbf{U}\mathbf{A}\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}\mathbf{A}'\mathbf{U}'\right]^{-1} \mathbf{U}\mathbf{A}\hat{\boldsymbol{\theta}},$$

$$Q_W(M_{LL}) = \left(\mathbf{U}_L\,\mathbf{A}\,\mathbf{ln}(\hat{\boldsymbol{\theta}})\right)' \left[\mathbf{U}\mathbf{A}\mathbf{D}^{-1}_{\hat{\boldsymbol{\theta}}}\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}\mathbf{D}^{-1}_{\hat{\boldsymbol{\theta}}}\mathbf{A}'\mathbf{U}'\right]^{-1} \mathbf{U}_L\,\mathbf{A}\,\mathbf{ln}(\hat{\boldsymbol{\theta}}),$$

em que $\mathbf{U}$ é ortogonal a $\mathbf{X}$ ($\mathbf{UX} = \mathbf{0}$) e $\mathbf{U}_L$ é ortogonal a $\mathbf{X}_L$ ($\mathbf{U}_L\mathbf{X}_L = \mathbf{0}$)

- Sob M (M_L ou M_{LL})

$$Q_V(M) \overset{a}{\approx} Q_P(M) \overset{a}{\approx} Q_N(M) \overset{a}{\approx} Q_W(M) \xrightarrow{a} \chi^2_{(u-p)},$$

em que $u - p$ é o número de restrições impostas pelo modelo

Mínimos Quadrados Generalizados (MQG)

- Sejam $\tilde{\boldsymbol{\theta}}$ e $\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}}$ estimadores consistentes de $\boldsymbol{\theta}$ e de $\mathbf{V}_{\tilde{\boldsymbol{\theta}}}$ (*e.g.*, $\hat{\boldsymbol{\theta}}$ e $\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}}$), e suponha que

$$\tilde{\boldsymbol{\theta}} \overset{a}{\sim} N_{SR}\left(\boldsymbol{\theta}, \tilde{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}}\right)$$

- Aplicando o método *delta*, tem-se que

$$\tilde{\mathbf{F}} \equiv \mathbf{F}\left(\tilde{\boldsymbol{\theta}}\right) \overset{a}{\sim} N_u\left(\mathbf{F}, \tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}\right),$$

em que $\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}} = \tilde{\mathbf{G}}\tilde{\mathbf{V}}_{\tilde{\boldsymbol{\theta}}}\tilde{\mathbf{G}}'$, com $\tilde{\mathbf{G}} \equiv \mathbf{G}\left(\tilde{\boldsymbol{\theta}}\right) = \left.\dfrac{\partial \mathbf{F}}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$

- Estimador de MQG de $\boldsymbol{\beta}$

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}^{-1}\tilde{\mathbf{F}}, \qquad \text{com} \qquad \tilde{\mathbf{V}}_{\tilde{\boldsymbol{\beta}}} = \left(\mathbf{X}'\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}^{-1}\mathbf{X}\right)^{-1}$$

- Estatística de Wald ($\mathbf{U}$ é ortogonal a $\mathbf{X}$, *i.e.*, $\mathbf{UX} = \mathbf{0}$)

$$Q_W(M_F) = \left(\mathbf{U}\tilde{\mathbf{F}}\right)'\left[\mathbf{U}\tilde{\mathbf{V}}_{\tilde{\mathbf{F}}}\mathbf{U}'\right]^{-1}\mathbf{U}\tilde{\mathbf{F}} \underset{M_F}{\overset{a}{\longrightarrow}} \chi^2_{(u-p)}$$

Mínimos Quadrados Generalizados (MQG)

- Em muitos casos o vetor de funções $\mathbf{F}(\boldsymbol{\theta})$ pode ser expresso como uma composição de funções lineares, logarítmicas, exponenciais e adição de constantes.
- Exemplos:

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbf{A}_1\boldsymbol{\theta} \Longrightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{A}_1$$

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbf{ln}(\boldsymbol{\theta}) \Longrightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}_{\boldsymbol{\theta}}^{-1}$$

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbf{exp}(\boldsymbol{\theta}) \Longrightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}_{\mathbf{exp}(\boldsymbol{\theta})}$$

$$\mathbf{F}(\boldsymbol{\theta}) = \boldsymbol{\pi}_1 + \boldsymbol{\theta} \Longrightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{I}_{SR}$$

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbf{A}_1\ \mathbf{ln}(\boldsymbol{\theta}) \Longrightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{A}_1\ \mathbf{D}_{\boldsymbol{\theta}}^{-1}$$

$$\mathbf{F}(\boldsymbol{\theta}) = \mathbf{exp}[\mathbf{A}_1\ \mathbf{ln}(\boldsymbol{\pi}_1 + \boldsymbol{\theta})] \Longrightarrow \mathbf{G}(\boldsymbol{\theta}) = \mathbf{D}_{\mathbf{exp}[\mathbf{A}_1\ \mathbf{ln}(\boldsymbol{\pi}_1+\boldsymbol{\theta})]}\ \mathbf{A}_1\ \mathbf{D}_{\boldsymbol{\pi}_1+\boldsymbol{\theta}}^{-1}$$

Modelos estruturais lineares

- Formulação de equações livres

$$M_L : \mathbf{A}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta}$$

- Formulação em termos de restrições

$$M_L : \mathbf{U}\mathbf{A}\boldsymbol{\theta} = \mathbf{0}_{u-p}$$

- $\mathbf{A}$ é uma matriz $u \times SR$ com posto $r(\mathbf{A}) = u \leq S(R-1)$
- $\mathbf{A}$ deve ser linearm.indep. da matriz definidora das restrições naturais

$$\left[\mathbf{I}_S \otimes \mathbf{1}'_R\right] \boldsymbol{\theta} = \mathbf{1}_S \qquad r(\mathbf{A}', \mathbf{I}_S \otimes \mathbf{1}_R) = u + S$$

- O padrão de `Catdata` é $\mathbf{A} = \mathbf{I}_S \otimes [\mathbf{I}_{R-1}, \mathbf{0}_{R-1}]$
- $\mathbf{X}$ é uma matriz $u \times p$ com posto $r(\mathbf{X}) = p \leq u$ que especifica o modelo
- $\mathbf{U}$ é uma matriz $(u-p) \times u$ com as restrições do modelo e suas linhas são ortogonais às colunas de $\mathbf{X}$, ou seja, $\mathbf{U}\mathbf{X} = \mathbf{0}_{(u-p),p}$
- Modelo de simetria, de homogeneidade marginal e modelo linear geral

Problema da intenção de voto

- Sondagens realizadas sobre as intenções de voto de 445 pessoas em duas entrevistas espaçadas de um mês
- Objetivo: avaliar se as mudanças na intenção de voto são iguais nos dois sentidos

1ª sondagem	2ª sondagem		
	partido A	partido B	indeciso
partido A	192	1	5
partido B	2	146	5
indeciso	11	12	71

$$S = 1 \quad \text{e} \quad R = 3 \times 3 = 9$$

1ª	partido A			partido B			indeciso		
2ª	A	B	ind.	A	B	ind.	A	B	ind.
	192	1	5	2	146	5	11	12	71

Paulino e Singer (2006), Exemplo 3.1 / 8.1 / 11.2

Problema da intenção de voto

1ª sondagem	2ª sondagem		
	partido A ($j=1$)	partido B ($j=2$)	indeciso ($j=3$)
partido A ($i=1$)	θ_{11}	θ_{12}	θ_{13}
partido B ($i=2$)	θ_{21}	θ_{22}	θ_{23}
indeciso ($i=3$)	θ_{31}	θ_{32}	θ_{33}

Objetivo: avaliar se as mudanças na intenção de voto são iguais nos dois sentidos, *i.e.*, **hipótese de simetria**

$$H : \theta_{ij} = \theta_{ji}, \qquad i < j$$

$$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{31}, \theta_{32}, \theta_{33})'$$

Problema da intenção de voto

$$H : \theta_{ij} = \theta_{ji}, \qquad i < j$$

$$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{31}, \theta_{32}, \theta_{33})'$$

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{A}\boldsymbol{\theta} = \begin{pmatrix} \theta_{12} \\ \theta_{21} \\ \theta_{13} \\ \theta_{31} \\ \theta_{23} \\ \theta_{32} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 \\ \beta_2 \\ \beta_2 \\ \beta_3 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \mathbf{X}\boldsymbol{\beta},$$

em que $\mathbf{X} = \mathbf{I}_3 \otimes \mathbf{1}_2$

Problema da intenção de voto

$$H : \theta_{ij} - \theta_{ji} = 0, \qquad i < j$$

$$\mathbf{A}\boldsymbol{\theta} = (\theta_{12}, \theta_{21}, \theta_{13}, \theta_{31}, \theta_{23}, \theta_{32})'$$

$$\mathbf{U} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

$$\mathbf{U}\mathbf{A}\boldsymbol{\theta} = \begin{pmatrix} \theta_{12} - \theta_{21} \\ \theta_{13} - \theta_{31} \\ \theta_{23} - \theta_{32} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \mathbf{0}_3$$

Note que $\mathbf{U} = \mathbf{I}_3 \otimes (1, -1)$

Estudo de suscetibilidade à cárie dentária

- 97 crianças de 11 a 13 anos de uma escola pública
- Objetivo: avaliar se o método simplificado (mais barato) é tão eficaz quanto o método convencional (de custos elevados)

Método	Método convencional		
simplificado	baixo	médio	alto
baixo	11	5	0
médio	14	34	7
alto	2	13	11

$$S = 1 \quad \text{e} \quad R = 3 \times 3 = 9$$

Simpl.	baixo			médio			alto		
Conv.	baixo	médio	alto	baixo	médio	alto	baixo	médio	alto
	11	5	0	14	34	7	2	13	11

Paulino e Singer (2006), Exemplo 1.2 / 3.2 / 8.2 / 10.2 / 11.3 / 11.12

Estudo de suscetibilidade à cárie dentária

Método simplificado	Método convencional			Total
	baixo ($j=1$)	médio ($j=2$)	alto ($j=3$)	
baixo ($i=1$)	θ_{11}	θ_{12}	θ_{13}	θ_{1+}
médio ($i=2$)	θ_{21}	θ_{22}	θ_{23}	θ_{2+}
alto ($i=3$)	θ_{31}	θ_{32}	θ_{33}	θ_{3+}
Total	θ_{+1}	θ_{+2}	θ_{+3}	1

Objetivo: avaliar se o método simplificado (mais barato) é "marginalmente" tão eficaz quanto o método convencional (de custos elevados), *i.e.*, **hipótese de homogeneidade marginal**

$$H : \theta_{i+} = \theta_{+i}, \qquad i = 1, 2$$

$$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{31}, \theta_{32}, \theta_{33})'$$

Estudo de suscetibilidade à cárie dentária

$$H : \theta_{i+} = \theta_{+i} = \beta_i, \qquad i = 1, 2$$

$$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{31}, \theta_{32}, \theta_{33})'$$

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$\mathbf{A}\boldsymbol{\theta} = \begin{pmatrix} \theta_{1+} \\ \theta_{2+} \\ \theta_{+1} \\ \theta_{+2} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \mathbf{X}\boldsymbol{\beta},$$

em que $\mathbf{X} = \mathbf{1}_2 \otimes \mathbf{I}_2$

Problema do tamanho da ninhada

- Estudo de fertilidade de ovelhas de vários rebanhos
- Objetivo: avaliar a influência da fazenda e da raça no tamanho da ninhada

Fazenda	Raça	Tamanho da ninhada				Total
		0	1	2	≥ 3	
1	A	10	21	96	23	150
	B	4	6	28	8	46
	C	9	7	58	7	81
2	A	8	19	44	1	72
	B	5	17	56	1	79
	C	1	5	20	2	28
3	A	22	95	103	4	224
	B	18	49	62	0	129
	C	4	12	16	2	34

$$S = 3 \times 3 = 9 \quad \text{e} \quad R = 4$$

Paulino e Singer (2006), Exemplo 1.9 / 3.3 / 8.3 / 11.4

Problema do tamanho da ninhada

$\theta_{k(ij)}$: probabilidade da fazenda i e da raça j ter o tamanho da ninhada k

Admitindo que ninhadas de tamanho maior que 3 são raras, pode-se comparar os tamanhos médios

$$\mu_{ij} = 0 \times \theta_{1(ij)} + 1 \times \theta_{2(ij)} + 2 \times \theta_{3(ij)} + 3 \times \theta_{4(ij)}$$

Considerando que

$$\boldsymbol{\theta} = \left(\theta_{1(11)}, \ldots, \theta_{4(11)}, \theta_{1(12)}, \ldots, \theta_{4(12)}, \ldots, \ldots, \theta_{1(33)}, \ldots, \theta_{4(33)}\right)'$$

e

$$\mathbf{A} = \mathbf{I}_9 \otimes (0, 1, 2, 3)$$

chega-se a

$$\mathbf{A}\boldsymbol{\theta} = \left(\mu_{11}, \mu_{12}, \ldots, \mu_{33}\right)'$$

Problema do tamanho da ninhada

$$\mathbf{A}\boldsymbol{\theta} = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \\ \mu_{31} \\ \mu_{32} \\ \mu_{33} \end{pmatrix} = \begin{pmatrix} \beta \\ \beta + \alpha_2 \\ \beta + \alpha_3 \\ \beta + \gamma_2 \\ \beta + \gamma_2 + \alpha_2 \\ \beta + \gamma_2 + \alpha_3 \\ \beta + \gamma_3 \\ \beta + \gamma_3 + \alpha_2 \\ \beta + \gamma_3 + \alpha_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta \\ \gamma_2 \\ \gamma_3 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \mathbf{X}\boldsymbol{\beta}$$

Pode-se resumir o modelo por

$$\mu_{ij} = \beta + \gamma_i + \alpha_j, \qquad i, j, = 1, 2, 3,$$

com restrições de identificabilidade $\gamma_1 = \alpha_1 = 0$

Problema do tamanho da ninhada

$$\mu_{ij} = \beta + \gamma_i + \alpha_j, \qquad i, j, = 1, 2, 3,$$

com restrições de identificabilidade $\gamma_1 = \alpha_1 = 0$

β: tamanho médio da ninhada para qualquer ovelha da fazenda 1 e raça A

γ_i: diferença entre o tamanho médio da ninhada das fazendas i e 1 para qualquer ovelha de qualquer raça

α_j: diferença entre o tamanho médio da ninhada das raças j e 1 para qualquer ovelha de qualquer fazenda

Modelos estruturais log-lineares

- Úteis na descrição de padrões de associação entre variáveis categorizadas
- Formulação de equações livres

$$M_{LL} : \mathbf{ln}(\boldsymbol{\theta}) = [\mathbf{I}_S \otimes \mathbf{1}_R]\,\boldsymbol{\nu} + \mathbf{X}\boldsymbol{\beta}$$

- Formulação em termos de restrições

$$M_{LL} : \mathbf{U}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_{S(R-1)-p}$$

- $\boldsymbol{\nu}$ é uma componente associada às restrições naturais
- $\mathbf{X} = \left(\mathbf{X}_1', \ldots, \mathbf{X}_S'\right)'$, de dimensão $SR \times p$, é tal que cada submatriz $\mathbf{X}_s$, de dimensão $R \times p$, tem suas colunas linearmente independentes do vetor $\mathbf{1}_R$ definidor da s-ésima restrição natural, $\mathbf{1}_R'\boldsymbol{\theta}_s = 1$, *i.e.*, $r(\mathbf{1}_R, \mathbf{X}_s) = 1 + r(\mathbf{X}_s)$, $s = 1, \ldots, S$ e $r(\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}) = S + p$, $p \leq S(R-1)$
- $\mathbf{U}$ possui dimensão $(S[R-1] - p) \times SR$, tal que $\mathbf{U}[\mathbf{I}_S \otimes \mathbf{1}_R, \mathbf{X}] = \mathbf{0}_{(SR-p),p}$

Modelos estruturais log-lineares

Pode-se também considerar uma classe mais ampla de modelos log-lineares

- Formulação de equações livres

$$M_{LL} : \mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{X}_L\boldsymbol{\beta}$$

- Formulação em termos de restrições

$$M_{LL} : \mathbf{U}_L\,\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_{u-p}$$

- $\mathbf{A}$ é uma matriz $u \times SR$ com posto $r(\mathbf{A}) = u \leq S(R-1)$
- As linhas de $\mathbf{A}$ devem ser ortogonais às colunas da matriz definidora das restrições naturais, *i.e.*, $\mathbf{A}\,(\mathbf{I}_S \otimes \mathbf{1}_R) = \mathbf{0}_{u,S}$
- O padrão de `Catdata` é $\mathbf{A} = \mathbf{I}_S \otimes [\mathbf{I}_{R-1}, -\mathbf{1}_{R-1}]$
  logitos de referência com relação à categoria R
- $\mathbf{X}_L$ é $u \times p$ com posto $r(\mathbf{X}) = p \leq u$ que especifica o modelo
- $\mathbf{U}_L$ possui dimensão $(u-p) \times u$, tal que $\mathbf{U}_L\mathbf{X}_L = \mathbf{0}_{(u-p),p}$

Problema da anemia

- Estudo da FSP–USP com 128 crianças com 4 meses de idade em região com raras situações de desnutrição e miséria extrema
- Objetivo: avaliar se o aleitamento materno e a anemia estão associadas

Anemia	Aleitamento	
	apenas materno	misto
sim	3	25
não	32	68

$$S = 1 \quad \text{e} \quad R = 2 \times 2 = 4$$

Anemia	sim		não	
Aleitamento	materno	misto	materno	misto
	3	25	32	68

Paulino e Singer (2006), Exemplo 9.1 / 11.5

Problema da anemia

Anemia	Aleitamento (B)	
(A)	materno ($j=1$)	misto ($j=2$)
sim ($i=1$)	θ_{11}	θ_{12}
não ($i=2$)	θ_{21}	θ_{22}

Objetivo: avaliar se o aleitamento materno (B) e a anemia (A) estão associadas, *i.e.*, **hipótese de independência** estocástica entre A e B

$$H : \theta_{ij} = \theta_{i+}\theta_{+j}, \quad i,j = 1,2$$

$$\frac{\theta_{11}/\theta_{12}}{\theta_{21}/\theta_{22}} = \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}} \stackrel{H}{=} \frac{\theta_{1+}\theta_{+1}\ \theta_{2+}\theta_{+2}}{\theta_{1+}\theta_{+2}\ \theta_{2+}\theta_{+1}} = 1$$

$$H : \ln \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}} = 0$$

Problema da anemia

$$H : \ln \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}} = 0$$

$$H : \ln \theta_{11} - \ln \theta_{12} - \ln \theta_{21} + \ln \theta_{22} = 0$$

$$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})'$$

$$\mathbf{U} = \mathbf{A} = (1, -1, -1, 1)$$

$$\mathbf{U}\ \mathbf{ln}(\boldsymbol{\theta}) = 0$$

$$\mathbf{X}_L = 1$$

$$\mathbf{A}\ \mathbf{ln}(\boldsymbol{\theta}) = \mathbf{X}_L\beta$$

$\exp(\beta) = \dfrac{\theta_{11}/\theta_{21}}{\theta_{12}/\theta_{22}}$ é a razão de chances (chance de um bebê de 4 meses que é alimentado apenas com leite materno ter *vs.* não ter anemia em relação à mesma chance para um bebê que não recebe apenas leite materno)

Problema da anemia

$$\mathbf{ln}(\boldsymbol{\theta}) = \nu + \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} u_2^A \\ u_2^B \\ u_{22}^{AB} \end{pmatrix}$$

$$\ln \theta_{11} = \nu$$

$$\ln \theta_{12} = \nu + u_2^B \longrightarrow u_2^B = \ln(\theta_{12}/\theta_{11})$$

$$\ln \theta_{21} = \nu + u_2^A \longrightarrow u_2^A = \ln(\theta_{21}/\theta_{11})$$

$$\ln \theta_{22} = \nu + u_2^A + u_2^B + u_{22}^{AB} = \ln \theta_{11} + \ln(\theta_{21}/\theta_{11}) + \ln(\theta_{12}/\theta_{11}) + u_{22}^{AB}$$

$$u_{22}^{AB} = \ln \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}}$$

$$\ln \theta_{ij} = \nu + u_i^A + u_j^B + u_{ij}^{AB}, \qquad i, j = 1, 2$$

com restrições de identificabilidade $u_1^A = u_1^B = u_{11}^{AB} = u_{12}^{AB} = u_{21}^{AB} = 0$

Estudo da satisfação com o emprego

- 96 homens dos EUA foram sondados
- Objetivo: avaliar se o salário influencia a satisfação com o emprego

Renda anual (US$)	Satisfação com o emprego			
	muito insatisfeito	um pouco insatisfeito	um pouco satisfeito	muito satisfeito
$<$15,000	1	3	10	6
15,000$-$25,000	2	3	10	7
25,000$-$40,000	1	6	14	12
$>$40,000	0	1	9	11

$$S = 1 \quad \text{e} \quad R = 4 \times 4 = 16$$

Renda	$<$15,000				$\cdots$	$>$40,000			
Satisf.	MI	PI	PS	MS	$\cdots$	MI	PI	PS	MS
	1	3	10	6	$\cdots$	0	1	9	11

Agresti (2002), Seção 2.4.3

Estudo da satisfação com o emprego

$$H : \theta_{ij} = \theta_{i+}\theta_{+j}, \quad i, j = 1, 2, 3, 4$$

$$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{14}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{24}, \theta_{31}, \theta_{32}, \theta_{33}, \theta_{34}, \theta_{41}, \theta_{42}, \theta_{43}, \theta_{44})'$$

$$\mathbf{U}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{0}_9$$

$$\mathbf{U} = \left(\begin{array}{cccc|cccc|cccc|cccc}
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1
\end{array}\right)$$

Estudo da satisfação com o emprego

$$\mathbf{A} = \left(\begin{array}{cccc|cccc|cccc|cccc}
1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1
\end{array}\right)$$

$$\mathbf{A}\,\ln(\boldsymbol{\theta}) = \boldsymbol{\omega} = \mathbf{X}_L\beta, \qquad \boldsymbol{\omega} = (\omega_{11}, \omega_{12}, \ldots, \omega_{33})', \qquad \mathbf{X}_L = \mathbf{1}_9$$

$$\omega_{ij} = \ln\left(\frac{\theta_{ij}\theta_{i+1,j+1}}{\theta_{i,j+1}\theta_{i+1,j}}\right) = \beta, \qquad i,j = 1,2,3$$

$$\frac{\theta_{ij}\ \theta_{i',j'}}{\theta_{i,j'}\ \theta_{i',j}} = \exp\{\beta(i'-i)(j'-j)\}, \qquad \forall i < i', j < j'$$

Modelo de associação uniforme ou linear por linear ($u_{ij}^{AB} = \beta(a_i - \bar{a})(b_j - \bar{b})$) com escores $\{a_i\}$ e $\{b_j\}$ unitariamente espaçados

Problema da fobia em alcoólatras

- Estudo realizado com 93 alcoólatras
- Objetivo: avaliar se a presença de fobia, o consumo diário de álcool e a situação profissional estão relacionadas

Situação profissional	Uso diário de álcool	Fobia	
		sim	não
sem emprego	sim	10	24
	não	6	12
com emprego	sim	13	17
	não	4	7

$$S = 1 \quad \text{e} \quad R = 2 \times 2 \times 2 = 8$$

Emprego	não				sim			
Uso diário	sim		não		sim		não	
Fobia	S	N	S	N	S	N	S	N
	10	24	6	12	13	17	4	7

Paulino e Singer (2006), Exemplo 1.11 / 9.4

Problema da fobia em alcoólatras

Situação profissional (A)	Uso diário de álcool (B)	Fobia (C) sim ($k=1$)	Fobia (C) não ($k=2$)
s/emprego ($i=1$)	sim ($j=1$)	θ_{111}	θ_{112}
	não ($j=2$)	θ_{121}	θ_{122}
c/emprego ($i=2$)	sim ($j=1$)	θ_{211}	θ_{212}
	não ($j=2$)	θ_{221}	θ_{222}

$$(ABC) \quad \ln \theta_{ijk} = \nu + u_i^A + u_j^B + u_k^C + u_{ij}^{AB} + u_{ik}^{AC} + u_{jk}^{BC} + u_{ijk}^{ABC}$$

$$(AB, AC, BC) \quad H_{NI} : \frac{\theta_{ijk}\,\theta_{i'j'k}}{\theta_{i'jk}\,\theta_{ij'k}} = \omega_{ij,i'j'}, \text{indep.de } k \longleftrightarrow u_{ijk}^{ABC} = 0$$

$$(AB, AC) \quad H_{IC} : \theta_{ijk} = \frac{\theta_{ij+}\theta_{i+k}}{\theta_{i++}} \longleftrightarrow u_{jk}^{BC} = u_{ijk}^{ABC} = 0$$

$$(AB, C) \quad H_{IP} : \theta_{ijk} = \theta_{ij+}\theta_{++k} \longleftrightarrow u_{ik}^{AC} = u_{jk}^{BC} = u_{ijk}^{ABC} = 0$$

$$(A, B, C) \quad H_I : \theta_{ijk} = \theta_{i++}\theta_{+j+}\theta_{++k} \longleftrightarrow u_{ij}^{AB} = u_{ik}^{AC} = u_{jk}^{BC} = u_{ijk}^{ABC} = 0$$

Problema de uso de fio dental

- 30 crianças de cada sexo e faixa etária (5-8 e 9-12 anos) foram selecionadas de uma escola da rede pública do município de SP
- Objetivo: avaliar se o sexo e a faixa etária influenciam a freqüência e a habilidade no uso do fio dental

Sexo	Faixa etária	Freqüência	Habilidade	
			inábil	razoável
masc.	5-8	insuficiente	19	5
		boa	4	2
	9-12	insuficiente	5	8
		boa	0	17
fem.	5-8	insuficiente	11	6
		boa	7	6
	9-12	insuficiente	2	5
		boa	1	22

$$S = 2 \times 2 = 4 \quad \text{e} \quad R = 2 \times 2 = 4$$

Paulino e Singer (2006), Exemplo 1.5 / 9.11

Problema de uso de fio dental

$$\mathbf{A} = \mathbf{I}_4 \otimes (1,-1,-1,1)\,, \qquad \boldsymbol{\theta} = \left(\theta_{11(11)}, \theta_{12(11)}, \theta_{21(11)}, \theta_{22(11)},\ \theta_{11(12)}, \ldots, \theta_{22(22)}\right)'$$

$$\mathbf{A}\,\ln(\boldsymbol{\theta}) = \begin{pmatrix} \omega_{11} \\ \omega_{12} \\ \omega_{21} \\ \omega_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta \\ \gamma_2 \\ \alpha_2 \end{pmatrix} = \mathbf{X}_L \boldsymbol{\beta}$$

$$\omega_{ij} = \beta + \gamma_i + \alpha_j, \qquad i,j = 1,2,$$

com restrições de identificabilidade $\gamma_1 = \alpha_1 = 0$

- $\exp(\beta)$: razão de chances* (RC) para meninos de 5-8 anos
- a RC p/meninas é $\exp(\gamma_2)$ vezes a RC p/meninos de qualquer fx.etária
- a RC para 9-12 anos é $\exp(\alpha_2)$ vezes a RC para 5-8 de qualquer sexo

*chance de uma criança ter habilidade razoável *vs.* ser inábil no uso do fio dental quando o usa com uma “boa” freqüência em relação à mesma chance quando usa o fio dental com freqüência “insuficiente”

Problema da complicação pulmonar

- 1162 pacientes tiveram os graus de complicação pulmonar pré e pós-operatório avaliados
- Objetivo: comparar os riscos de complicação pulmonar do período pós-operatório entre os níveis da avaliação do pré-operatório

Avaliação de complicação pulmonar

Pré-	Pós-operatório	
-operatório	sem complicação	com complicação
baixo	737	48
moderado	243	74
alto	39	21

$$S = 3 \quad \text{e} \quad R = 2$$

Paulino e Singer (2006), Exemplo 6.5 / 10.4 / 11.10

Problema da complicação pulmonar

Avaliação de complicação pulmonar

Pré-	Pós-operatório	
-operatório	sem complicação ($j=1$)	com complicação ($j=2$)
baixo ($i=1$)	$\theta_{1(1)}$	$\theta_{2(1)}$
moderado ($i=2$)	$\theta_{1(2)}$	$\theta_{2(2)}$
alto ($i=3$)	$\theta_{1(3)}$	$\theta_{2(3)}$

Objetivo: comparar os riscos de complicação pulmonar do período pós-operatório entre os níveis da avaliação do pré-operatório, *e.g.*, **hipótese de igualdade de riscos relativos**

$$H: \frac{\theta_{2(2)}}{\theta_{2(1)}} = \frac{\theta_{2(3)}}{\theta_{2(2)}} = \exp(\beta)$$

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\theta} = \left(\theta_{1(1)}, \theta_{2(1)}, \theta_{1(2)}, \theta_{2(2)}, \theta_{1(3)}, \theta_{2(3)}\right)'$$

$$\mathbf{A}\,\mathbf{ln}(\boldsymbol{\theta}) = \mathbf{1}_2\beta = \mathbf{X}\boldsymbol{\beta}$$

(note que $\mathbf{A}$ não é ortogonal à matriz definidora das restrições naturais $\mathbf{I}_3 \otimes \mathbf{1}_2$)

Estudo de suscetibilidade à cárie dentária

- 97 crianças de 11 a 13 anos de uma escola pública
- Objetivo: avaliar se o método simplificado (mais barato) é tão eficaz quanto o método convencional (de custos elevados)

Método	Método convencional		
simplificado	baixo	médio	alto
baixo	11	5	0
médio	14	34	7
alto	2	13	11

$$S = 1 \quad \text{e} \quad R = 3 \times 3 = 9$$

Simpl.	baixo			médio			alto		
Conv.	baixo	médio	alto	baixo	médio	alto	baixo	médio	alto
	11	5	0	14	34	7	2	13	11

Paulino e Singer (2006), Exemplo 1.2 / 3.2 / 8.2 / 10.2 / 11.3 / 11.12

Estudo de suscetibilidade à cárie dentária

Para avaliar a concordância entre os dois métodos pode-se utilizar a medida *kappa* de Cohen

$$\kappa = \frac{\sum_{i=1}^{3} \theta_{ii} - \sum_{i=1}^{3} \theta_{i+}\theta_{+i}}{1 - \sum_{i=1}^{3} \theta_{i+}\theta_{+i}}$$

$$\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{13}, \theta_{21}, \theta_{22}, \theta_{23}, \theta_{31}, \theta_{32}, \theta_{33})'$$

$$\mathbf{A}_1 = \left(\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right) \longrightarrow \mathbf{A}_1\boldsymbol{\theta} = \begin{pmatrix} \theta_{11} + \theta_{22} + \theta_{33} \\ 1 \\ \theta_{1+} \\ \theta_{2+} \\ \theta_{3+} \\ \theta_{+1} \\ \theta_{+2} \\ \theta_{+3} \end{pmatrix}$$

Estudo de suscetibilidade à cárie dentária

$$\kappa = \frac{\sum_{i=1}^{3} \theta_{ii} - \sum_{i=1}^{3} \theta_{i+}\theta_{+i}}{1 - \sum_{i=1}^{3} \theta_{i+}\theta_{+i}}$$

$$\mathbf{A}_1\boldsymbol{\theta} = \left(\textstyle\sum_{i=1}^{3} \theta_{ii}, 1, \theta_{1+}, \theta_{2+}, \theta_{3+}, \theta_{+1}, \theta_{+2}, \theta_{+3}\right)'$$

$$\mathbf{A}_2 = \left(\begin{array}{cc|cccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array}\right) \rightarrow \mathbf{exp}[\mathbf{A}_2\ \mathbf{ln}(\mathbf{A}_1\boldsymbol{\theta})] = \left(\begin{array}{c} \sum_{i=1}^{3} \theta_{ii} \\ 1 \\ \theta_{1+}\theta_{+1} \\ \theta_{2+}\theta_{+2} \\ \theta_{3+}\theta_{+3} \end{array}\right)$$

$$\mathbf{A}_3 = \left(\begin{array}{c|c|ccc} 1 & 1 & -2 & -2 & -2 \\ 0 & 1 & -1 & -1 & -1 \end{array}\right)$$

$$\mathbf{A}_3\ \mathbf{exp}[\mathbf{A}_2\ \mathbf{ln}(\mathbf{A}_1\boldsymbol{\theta})] = \left(\begin{array}{c} \sum_{i=1}^{3} \theta_{ii} + 1 - 2\sum_{i=1}^{3} \theta_{i+}\theta_{+i} \\ 1 - \sum_{i=1}^{3} \theta_{i+}\theta_{+i} \end{array}\right)$$

Estudo de suscetibilidade à cárie dentária

$$\kappa = \frac{\sum_{i=1}^{3} \theta_{ii} - \sum_{i=1}^{3} \theta_{i+}\theta_{+i}}{1 - \sum_{i=1}^{3} \theta_{i+}\theta_{+i}}$$

$$\mathbf{A}_3 \, \mathbf{exp}[\mathbf{A}_2 \, \mathbf{ln}(\mathbf{A}_1\boldsymbol{\theta})] = \begin{pmatrix} \sum_{i=1}^{3} \theta_{ii} + 1 - 2\sum_{i=1}^{3} \theta_{i+}\theta_{+i} \\ 1 - \sum_{i=1}^{3} \theta_{i+}\theta_{+i} \end{pmatrix} \qquad \mathbf{A}_4 = (1, -1)$$

$$\mathbf{exp}(\mathbf{A}_4 \, \mathbf{ln}\{\mathbf{A}_3 \, \mathbf{exp}[\mathbf{A}_2 \, \mathbf{ln}(\mathbf{A}_1\boldsymbol{\theta})]\}) = \frac{\sum_{i=1}^{3} \pi_{ii} + 1 - 2\sum_{i=1}^{3} \pi_{i+}\pi_{+i}}{1 - \sum_{i=1}^{3} \pi_{i+}\pi_{+i}} = \kappa + 1$$

$$\boldsymbol{\pi}_1 = -1 \longrightarrow \boldsymbol{\pi}_1 + \mathbf{exp}(\mathbf{A}_4 \, \mathbf{ln}\{\mathbf{A}_3 \, \mathbf{exp}[\mathbf{A}_2 \, \mathbf{ln}(\mathbf{A}_1\boldsymbol{\theta})]\}) = \kappa$$

Tipos de dados com omissão

É comum encontrar problemas em que

- algumas pessoas
  - não cumpriram uma ou mais diretrizes do protocolo do estudo (*non-compliance*)
  - abandonaram o estudo (*drop-out*) durante sua realização
  - não responderam a certas questões (*non-response*)
- uma parte do banco de dados está faltando (*missing*) por um motivo qualquer

Nestes casos, **as respostas em algumas variáveis para uma parte das unidades experimentais não são observadas** e então, diz-se que o conjunto de dados obtido tem **omissão**.

Exemplo (Baker *et al.*, 1992)

- Estudo prospectivo do Departamento de Saúde dos Estados Unidos
- Mães grávidas (fumantes ou não) foram acompanhadas até ao parto
- Os recém-nascidos tiveram seu pesos classificados (< 2.5 kg ou ≥ 2.5 kg)
- Objetivo: avaliar a associação entre o hábito de fumo da mãe e o peso do recém-nascido

Mãe	Peso do recém-nascido (kg)		
fumante	< 2.5	≥ 2.5	omisso
sim	4 512	21 009	*1 049*
não	3 394	24 132	*1 135*
omisso	142	464	1 224

1% não têm o hábito de fumo da mãe

De 57 061 mães/recém-nascidos,

93% foram completamente categorizados

4% não têm o peso do recém-nascido

2% as duas informações estão faltando

Exemplo (cont.)

Mãe	Peso do recém-nascido (kg)		
fumante	< 2.5	≥ 2.5	omisso
sim	4 512	21 009	*1 049*
não	3 394	24 132	*1 135*
omisso	142	464	1 224

1% não têm o hábito de fumo da mãe

De 57 061 mães/recém-nascidos,

93% foram completamente categorizados

4% não têm o peso do recém-nascido

2% as duas informações estão faltando

Algumas especulações possíveis para as omissões são:

1. Problemas no armazenamento dos dados
2. Mães fumantes se negariam a informar o hábito de fumo mais do que as não-fumantes
3. Recém-nascidos com pesos menores poderiam ter complicações que impediriam sua pesagem no instante desejado

Análise de Casos Completos (ACC)

Mãe	Peso do recém-nascido (kg)		
fumante	< 2.5	≥ 2.5	omisso
sim	4 512	21 009	~~*1 049*~~
não	3 394	24 132	~~*1 135*~~
omisso	~~142~~	~~464~~	~~1 224~~

- abordagem simples e fácil
- consiste em ignorar os dados com omissão e analisar apenas os dados completamente categorizados
- supondo que constituem uma amostra aleatória da população de interesse
- implica, em geral, perda de eficiência dos estimadores e/ou vieses nas inferências de interesse, dependendo do tipo de omissão

Duas variáveis dicotômicas

Y_1	Y_2	
	1	2
1	θ_{11}	θ_{12}
2	θ_{21}	θ_{22}

$$\theta_{ij} = P(Y_1 = i, Y_2 = j)$$

$$\sum_{i=1}^{2}\sum_{j=1}^{2} \theta_{ij} = 1$$

Y_1	Y_2		
	1	2	omisso
1 2	$W = 1 W = 1$ (Y_1^o, Y_2^o)		$W = 2 W = 2$ (Y_1^o, Y_2^m)
omisso	$W = 3 W = 3$ (Y_1^m, Y_2^o)		$W = 4 W = 4$ (Y_1^m, Y_2^m)

W: representativa dos diferentes padrões de omissão

Modelo probabilístico

$$\gamma_{tij} = P(W = t, Y_1 = i, Y_2 = j) \qquad \sum_{t=1}^{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\gamma_{tij} = 1$$

	$W=1\ (Y_1^o, Y_2^o)$		$W=2\ (Y_1^o, Y_2^m)$		$W=3\ (Y_1^m, Y_2^o)$		$W=4\ (Y_1^m, Y_2^m)$	
$Y_1 \setminus Y_2$	1	2	1	2	1	2	1	2
1	γ_{111}	γ_{112}	γ_{211}	γ_{212}	γ_{311}	γ_{312}	γ_{411}	γ_{412}
2	γ_{121}	γ_{122}	γ_{221}	γ_{222}	γ_{321}	γ_{322}	γ_{421}	γ_{422}

n_{211}

n_{212} $\qquad \gamma_{211} + \gamma_{212} = \gamma_{21+}$

	$W=1\ (Y_1^o, Y_2^o)$		$W=2\ (Y_1^o, Y_2^m)$			$W=3\ (Y_1^m, Y_2^o)$		$W=4\ (Y_1^m, Y_2^m)$		
$Y_1 \setminus Y_2$	1	2	1	2	total	1	2	1	2	total
1	n_{111}	n_{112}	?	?	n_{21+}	?	?	?	?	
2	n_{121}	n_{122}	?	?	n_{22+}	?	?	?	?	
total						n_{3+1}	n_{3+2}			n_{4++}

? representa uma freqüência não-observável

Fatorações

$$\gamma_{tij} = P(W = t, Y_1 = i, Y_2 = j)$$
$$= P(W = t)P(Y_1 = i, Y_2 = j|W = t) = \phi_t \eta_{ij(t)}$$

Fatoração de **modelos de mistura de padrões** (*pattern-mixture models*)

$$P(Y_1 = i, Y_2 = j) = \theta_{ij} = \sum_{t=1}^{T} \phi_t \eta_{ij(t)}$$

Fatorações

$$\gamma_{tij} = P(W = t, Y_1 = i, Y_2 = j)$$
$$= P(Y_1 = i, Y_2 = j)P(W = t|Y_1 = i, Y_2 = j) = \theta_{ij}\lambda_{t(ij)}$$

um modelo marginal para as características de interesse

é combinado com

um modelo condicional para o mecanismo de omissão dado as características

$\{\theta_{ij}\}$: probabilidades marginais de categorização

$\{\lambda_{t(ij)}\}$: *probabilidades condicionais de omissão*

Fatoração de **modelos de seleção** (*selection models*)

Função de verossimilhança

$$\mathbf{N} = (n_{111}, n_{112}, n_{121}, n_{122}, n_{21+}, n_{22+}, n_{3+1}, n_{3+2}, n_{4++})' \sim \text{Multinomial}$$

$$L\left(\{\theta_{ij}\}, \{\lambda_{t(ij)}\} \middle| \mathbf{N}\right) \propto \overbrace{\prod_{i=1}^{2}\prod_{j=1}^{2}\left(\theta_{ij}\lambda_{1(ij)}\right)^{n_{1ij}}}^{W=1\ (Y_1^o, Y_2^o)} \times \overbrace{\prod_{i=1}^{2}\left(\theta_{i1}\lambda_{2(i1)} + \theta_{i2}\lambda_{2(i2)}\right)^{n_{2i+}}}^{W=2\ (Y_1^o, Y_2^m)} \times$$

$$\underbrace{\prod_{j=1}^{2}\left(\theta_{1j}\lambda_{3(1j)} + \theta_{2j}\lambda_{3(2j)}\right)^{n_{3+j}}}_{W=3\ (Y_1^m, Y_2^o)} \times \underbrace{\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}\lambda_{4(ij)}\right)^{n_{4++}}}_{W=4\ (Y_1^m, Y_2^m)}$$

com $\displaystyle\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij} = 1$ e $\displaystyle\sum_{t=1}^{4}\lambda_{t(ij)} = 1, \quad i, j = 1, 2$

Mecanismo de omissão: MAR

- *Missing At Random* (**MAR**), omissão aleatória ou omissão não-informativa
- **As probabilidades condicionais de omissão dependem apenas do que é observado**

$$\lambda_{1(ij)} = \alpha_{1(ij)} \qquad \lambda_{2(ij)} = \alpha_{2(i)} \qquad \lambda_{3(ij)} = \alpha_{3(j)} \qquad \lambda_{4(ij)} = \alpha_4$$

	$W = 1\,(Y_1^o, Y_2^o)$		$W = 2\,(Y_1^o, Y_2^m)$		$W = 3\,(Y_1^m, Y_2^o)$		$W = 4\,(Y_1^m, Y_2^m)$	
$Y_1 \setminus Y_2$	1	2	1	2	1	2	1	2
1	$\alpha_{1(11)}$	$\alpha_{1(12)}$	$\alpha_{2(1)}$	$\alpha_{2(1)}$	$\alpha_{3(1)}$	$\alpha_{3(2)}$	α_4	α_4
2	$\alpha_{1(21)}$	$\alpha_{1(22)}$	$\alpha_{2(2)}$	$\alpha_{2(2)}$	$\alpha_{3(1)}$	$\alpha_{3(2)}$	α_4	α_4

$$\alpha_{1(ij)} = 1 - \alpha_{2(i)} - \alpha_{3(j)} - \alpha_4$$

Distrib.valores omissos sob MAR

	$W=1\ (Y_1^o, Y_2^o)$		$W=2\ (Y_1^o, Y_2^m)$			$W=3\ (Y_1^m, Y_2^o)$		$W=4\ (Y_1^m, Y_2^m)$		
$Y_1 \setminus Y_2$	1	2	1	2	total	1	2	1	2	total
1	n_{111}	n_{112}	?	?	n_{21+}	?	?	?	?	
2	n_{121}	n_{122}	?	?	n_{22+}	?	?	?	?	
total						n_{3+1}	n_{3+2}			n_{4++}

? representa uma freqüência não-observável

$$P(Y_2=1|Y_1=1,W=2) = \frac{P(W=2,Y_1=1,Y_2=1)}{P(W=2,Y_1=1)} = \frac{\gamma_{211}}{\gamma_{21+}} = \frac{\gamma_{211}}{\gamma_{211}+\gamma_{212}}$$

$$= \frac{\theta_{11}\alpha_{2(1)}}{\theta_{11}\alpha_{2(1)}+\theta_{12}\alpha_{2(1)}} = \frac{\theta_{11}}{\theta_{11}+\theta_{12}} = \frac{\theta_{11}}{\theta_{1+}}$$

$$= P(Y_2=1|Y_1=1)$$

Condicionalmente ao que foi observado, as freqüências com um resultado omisso estariam **distribuídas entre as categorias omissas como se não houvesse omissão**

ACC sob MAR

$$P(Y_1 = i, Y_2 = j|W = 1) = \frac{P(W = 1, Y_1 = i, Y_2 = j)}{P(W = 1)} = \frac{\gamma_{1ij}}{\gamma_{1++}} = \frac{\gamma_{1ij}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\gamma_{1ij}}$$

$$= \frac{\theta_{ij}\alpha_{1(ij)}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}\alpha_{1(ij)}} = \frac{\theta_{ij}(1 - \alpha_{2(i)} - \alpha_{3(j)} - \alpha_4)}{\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}(1 - \alpha_{2(i)} - \alpha_{3(j)} - \alpha_4)}$$

$$= \theta_{ij}\left(\frac{1 - \alpha_{2(i)} - \alpha_{3(j)} - \alpha_4}{1 - \alpha_{2(1)}\theta_{1+} - \alpha_{2(2)}\theta_{2+} - \alpha_{3(1)}\theta_{+1} - \alpha_{3(2)}\theta_{+2} - \alpha_4}\right)$$

Uma ACC sob o mecanismo MAR

leva, em geral, a

inferências enviesadas sobre θ

Valores de $P(Y_1 = i, Y_2 = j|W = 1)$

δ	$\{\theta_{ij}\}$	α		
		0.3	0.5	0.7
-0.2	0.1	0.0676	0.0556	0.0294
	0.3	0.2027	0.1667	0.08820.0882
	0.4	0.4865	0.5185	0.58820.5882
	0.2	0.2432	0.2593	0.2941
-0.1	0.1	0.0833	0.0769	0.0625
	0.30.3	0.2500	0.2308	0.1875
	0.40.4	0.4444	0.4615	0.5000
	0.2	0.2222	0.2308	0.2500
0.1	0.1	0.1176	0.1250	0.1429
	0.3	0.35290.3529	0.3750	0.4286
	0.4	0.35290.3529	0.3333	0.2857
	0.2	0.1765	0.1667	0.1429
0.2	0.1	0.1364	0.1522	0.1923
	0.3	0.4091	0.4565	0.57690.5769
	0.4	0.3030	0.2609	0.15380.1538
	0.2	0.1515	0.1304	0.0769

Suponha apenas $W = 1, 2$

sob o mecanismo MAR, co

$(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$
$= (0.1, 0.3, 0.4, 0.2)$,

$\alpha_{2(1)} = P(Y_2^m|Y_1 = 1) = \alpha$

e

$\alpha_{2(2)} = P(Y_2^m|Y_1 = 2) = \alpha$

Avaliação da simetria:

$\theta_{21} - \theta_{12} = 0.4 - 0.3 = 0.1$

$P(Y_1 = 2, Y_2 = 1|W = 1)$
$-P(Y_1 = 1, Y_2 = 2|W = 1)$

$= 0.5882 - 0.0882 = 0.500$

$= 0.3529 - 0.3529 = 0.000$

$= 0.1538 - 0.5769 = -0.42$

Função de verossimilhança sob MAR

$$L\left(\boldsymbol{\theta},\boldsymbol{\alpha}|\,\mathbf{N};\mathrm{MAR}\right)\propto\prod_{i=1}^{2}\prod_{j=1}^{2}\left(\theta_{ij}\alpha_{1(ij)}\right)^{n_{1ij}}\prod_{i=1}^{2}\left(\theta_{i1}\alpha_{2(i)}+\theta_{i2}\alpha_{2(i)}\right)^{n_{2i+}}\times$$

$$\prod_{j=1}^{2}\left(\theta_{1j}\alpha_{3(j)}+\theta_{2j}\alpha_{3(j)}\right)^{n_{3+j}}\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}\alpha_{4}\right)^{n_{4++}}$$

$$=\prod_{i=1}^{2}\prod_{j=1}^{2}\theta_{ij}^{n_{1ij}}\prod_{i=1}^{2}\left(\theta_{i1}+\theta_{i2}\right)^{n_{2i+}}\prod_{j=1}^{2}\left(\theta_{1j}+\theta_{2j}\right)^{n_{3+j}}\times$$

$$\prod_{i=1}^{2}\prod_{j=1}^{2}\alpha_{1(ij)}^{n_{1ij}}\prod_{i=1}^{2}\alpha_{2(i)}^{n_{2i+}}\prod_{j=1}^{2}\alpha_{3(j)}^{n_{3+j}}\alpha_{4}^{n_{4++}}$$

$$\equiv L_1\left(\boldsymbol{\theta}\,|\mathbf{N}\,\right)L_2\left(\boldsymbol{\alpha}\,|\mathbf{N};\mathrm{MAR}\right)$$

Inferências sob MAR

$$L\left(\boldsymbol{\theta},\boldsymbol{\alpha}|\,\mathbf{N};\mathrm{MAR}\right)\propto\prod_{i=1}^{2}\prod_{j=1}^{2}\theta_{ij}^{n_{1ij}}\prod_{i=1}^{2}\left(\theta_{i1}+\theta_{i2}\right)^{n_{2i+}}\prod_{j=1}^{2}\left(\theta_{1j}+\theta_{2j}\right)^{n_{3+j}}\times$$

$$\prod_{i=1}^{2}\prod_{j=1}^{2}\alpha_{1(ij)}^{n_{1ij}}\prod_{i=1}^{2}\alpha_{2(i)}^{n_{2i+}}\prod_{j=1}^{2}\alpha_{3(j)}^{n_{3+j}}\alpha_{4}^{n_{4++}}$$

$$\equiv L_1\left(\boldsymbol{\theta}\,|\mathbf{N}\,\right)L_2\left(\boldsymbol{\alpha}\,|\mathbf{N};\mathrm{MAR}\,\right),$$

- n_{4++} não traz informação na estimação de $\boldsymbol{\theta}$
- se $\boldsymbol{\theta}$ e $\boldsymbol{\alpha}$ forem funcionalmente independentes
  - $\boldsymbol{\theta}$ e $\boldsymbol{\alpha}$ podem ser estimados separadamente
  - $\mathrm{IO}_1\left(\boldsymbol{\theta}\right)$ não depende de $\boldsymbol{\alpha}$, e sob o ponto de vista de inferência sobre $\boldsymbol{\theta}$ baseada apenas na verossimilhança, o processo de omissão **é** ignorável
  - $\mathrm{E}(n_{tij})=n_{+++}\theta_{ij}\lambda_{t(ij)}\Longrightarrow\mathrm{IF}_1\left(\boldsymbol{\theta},\boldsymbol{\alpha};\mathrm{MAR}\right)$ depende de $\boldsymbol{\alpha}$, logo, sob o ponto de vista de inferências freqüentistas sobre $\boldsymbol{\theta}$, o processo de omissão **não é** ignorável

Estimação de θ sob MAR

	$W=1\ (Y_1^o, Y_2^o)$		$W=2\ (Y_1^o, Y_2^m)$			$W=3\ (Y_1^m, Y_2^o)$		$W=4\ (Y_1^m, Y_2^m)$		
$Y_1 \setminus Y_2$	1	2	1	2	total	1	2	1	2	total
1	n_{111}	n_{112}	?	?	n_{21+}	?	?	?	?	
2	n_{121}	n_{122}	?	?	n_{22+}	?	?	?	?	
total						n_{3+1}	n_{3+2}			n_{4++}

? representa uma freqüência não-observável

$$\theta_{ij} = P(Y_1 = i, Y_2 = j) = P(Y_1 = i)P(Y_2 = j|Y_1 = i) = \theta_{i+}\theta_{j(i)}$$

$$\hat{\theta}_{i+} = \frac{n_{1i+} + n_{2i+}}{n_{1++} + n_{2++}}, \quad i = 1, 2, \qquad \hat{\theta}_{j(i)} = \frac{n_{1ij}}{n_{1i+}}, \quad j, i = 1, 2$$

Mecanismo de omissão: MCAR

- *Missing Completely At Random* (**MCAR**) ou omissão completamente aleatória
- **As probabilidades condicionais de omissão independem do que é e do que não é observado**

$$\lambda_{t(ij)} = \alpha_t, \qquad t = 1, 2, 3, 4$$

	$W = 1\ (Y_1^o, Y_2^o)$		$W = 2\ (Y_1^o, Y_2^m)$		$W = 3\ (Y_1^m, Y_2^o)$		$W = 4\ (Y_1^m, Y_2^m)$	
$Y_1 \setminus Y_2$	1	2	1	2	1	2	1	2
1	α_1	α_1	α_2	α_2	α_3	α_3	α_4	α_4
2	α_1	α_1	α_2	α_2	α_3	α_3	α_4	α_4

$\alpha_1 = 1 - \alpha_2 - \alpha_3 - \alpha_4$

Inferências sob MCAR

$$L\left(\boldsymbol{\theta},\boldsymbol{\alpha}|\,\mathbf{N};\mathrm{MCAR}\right)\propto\prod_{i=1}^{2}\prod_{j=1}^{2}\theta_{ij}^{n_{1ij}}\prod_{i=1}^{2}\left(\theta_{i1}+\theta_{i2}\right)^{n_{2i+}}\prod_{j=1}^{2}\left(\theta_{1j}+\theta_{2j}\right)^{n_{3+j}}\times$$

$$\prod_{t=1}^{4}\alpha_t^{n_{t++}}$$

$$\equiv L_1\left(\boldsymbol{\theta}\,|\mathbf{N}\right)L_2\left(\boldsymbol{\alpha}\,|\mathbf{N}_{++};\mathrm{MCAR}\right),$$

$$\mathbf{N}_{++}=(n_{1++},n_{2++},n_{3++},n_{4++})'$$

- $L_1\left(\boldsymbol{\theta}\,|\mathbf{N}\right)$ é igual sob os mecanismos MAR e MCAR
- $\mathbf{N}_{++}$ é uma estatística suficiente específica para $\boldsymbol{\alpha}$ e ancilar específica para $\boldsymbol{\theta}$
- Portanto, pode-se analisar os dados através de $L_1\left(\boldsymbol{\theta}\,|\mathbf{N}\right)$, núcleo de uma distribuição produto de multinomiais $M_4(n_{1++},\boldsymbol{\theta})$, $Bin(n_{2++},\theta_{1+})$ e $Bin(n_{3++},\theta_{+1})$

ACC sob MCAR

$$P(Y_1 = i, Y_2 = j|W = 1) = \frac{P(W = 1, Y_1 = i, Y_2 = j)}{P(W = 1)} = \frac{\gamma_{1ij}}{\gamma_{1++}}$$
$$= \frac{\gamma_{1ij}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\gamma_{1ij}} = \frac{\theta_{ij}\alpha_1}{\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}\alpha_1} = \theta_{ij}$$
$$= P(Y_1 = i, Y_2 = j)$$

Uma ACC sob o mecanismo MCAR

não leva a inferências

enviesadas sobre θ

Ganho de informação sob MCAR

$$\left(\frac{\partial^2 L_1(\boldsymbol{\theta}\,|\mathbf{N})}{\partial\theta_{11}^2}\right) = n_{1++}\left[\frac{1}{\theta_{11}} + \frac{1}{1-\theta_{11}-\theta_{12}-\theta_{21}}\right] + +$$

$$n_{2++}\left[\frac{1}{\theta_{11}+\theta_{12}} + \frac{1-\theta_{11}-\theta_{12}}{(1-\theta_{11}+\theta_{12})^2}\right] + n_{2++}\left[\frac{1}{\theta_{11}+\theta_{12}} + \frac{1-\theta_{11}-\theta_{12}}{(1-\theta_{11}+\theta_{12}}\right.$$

$$n_{3++}\left[\frac{1}{\theta_{11}+\theta_{21}} + \frac{1-\theta_{11}-\theta_{21}}{(1-\theta_{11}+\theta_{21})^2}\right] n_{3++}\left[\frac{1}{\theta_{11}+\theta_{21}} + \frac{1-\theta_{11}-\theta_{21}}{(1-\theta_{11}+\theta_{21})^2}\right]$$

obtido na ACC

termos positivos

Ganho de informação em θ_{11} sob MCAR

Estudo de simulação

- $(\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}) = (0.1, 0.4, 0.3, 0.2)$
- Apenas 2 cenários de omissão sob o MCAR ($W = 1, 2$)

	$W = 1\ (Y_1^o, Y_2^o)$		$W = 2\ (Y_1^o, Y_2^m)$	
$Y_1 \setminus Y_2$	1	2	1	2
1	$1 - \alpha_2$	$1 - \alpha_2$	α_2	α_2
2	$1 - \alpha_2$	$1 - \alpha_2$	α_2	α_2

- $\alpha_2 = 0.2,\ 0.5,\ 0.8$
- $n_{+++} = 10,\ 20,\ 50,\ 100,\ 200,\ 500$
- 1 milhão de réplicas de Monte Carlo
- Desvio padrão das EMV sob a ACC e o mecanismo MCAR
- "Ganho": $\dfrac{DP(\text{ACC}) - DP(\text{MCAR})}{DP(\text{ACC})}$

n_{+++}	$\{\theta_{ij}\}$	$\alpha_2 = 0.2$	0.5	0.8
10	0.1	-0.2% -0.2%	2.8% 2.8%	22.5% 22.5%
	0.4	6.9%	19.6% 19.6%	40.7% 40.7%
	0.3	3.8%	13.2% 13.2%	35.7% 35.7%
	0.2	1.6%	8.1% 8.1%	30.8% 30.8%
20	0.1	0.4%	-0.3% -0.3%	10.1% 10.1%
	0.4	6.9%	17.8%	35.5% 35.5%
	0.3	4.1%	9.7%	26.3% 26.3%
	0.2	2.0%	4.0%	18.8% 18.8%
50	0.1	0.9%	1.8%	-0.9% -0.9%
	0.4	7.0%	18.3%	29.5%
	0.3	4.3%	10.8%	15.4%
	0.2	2.3%	5.6%	6.2%
100	0.1	1.0%	2.3%	2.1%
	0.4	6.6%	18.3%	30.8%
	0.3	4.3%	11.1%	17.3%
	0.2	2.5%	6.2%	8.4%
200	0.1	1.0%	2.6%	3.5%
	0.4	6.9%	18.3%	31.3%
	0.3	4.4%	11.2%	18.3%
	0.2	2.5%	6.3%	9.6%

Ganhos exagerado
se $E(n_{1ij}) \leq 1$
para dois ou mais (

Ganhos negativos
onde $E(n_{111}) \leq 1$ e
$E(n_{1ij}) \geq 1.5,\ (i,j)$

Quando $E(n_{1ij}) \geq$
para todos (i,j),
o ganho é maior co

α_2 aumenta

$\{\theta_{ij}\} \to 0.5$

n_{+++} aumenta

Mecanismo de omissão: MNAR

- *Missing Not At Random* (**MNAR**), omissão não-aleatória ou omissão informativa
- **As probabilidades condicionais de omissão dependem de algum modo do que não foi observado**

	$W=1\ (Y_1^o, Y_2^o)$		$W=2\ (Y_1^o, Y_2^m)$		$W=3\ (Y_1^m, Y_2^o)$		$W=4\ (Y_1^m, Y_2^m)$	
$Y_1 \setminus Y_2$	1	2	1	2	1	2	1	2
1	γ_{111}	γ_{112}	γ_{211}	γ_{212}	γ_{311}	γ_{312}	γ_{411}	γ_{412}
2	γ_{121}	γ_{122}	γ_{221}	γ_{222}	γ_{321}	γ_{322}	γ_{421}	γ_{422}

	$W=1\ (Y_1^o, Y_2^o)$		$W=2\ (Y_1^o, Y_2^m)$			$W=3\ (Y_1^m, Y_2^o)$		$W=4\ (Y_1^m, Y_2^m)$		
$Y_1 \setminus Y_2$	1	2	1	2	total	1	2	1	2	total
1	n_{111}	n_{112}	?	?	n_{21+}	?	?	?	?	
2	n_{121}	n_{122}	?	?	n_{22+}	?	?	?	?	
total						n_{3+1}	n_{3+2}			n_{4++}

? representa uma freqüência não-observável

Inferências sob MNAR

	$W=1\,(Y_1^o, Y_2^o)$		$W=2\,(Y_1^o, Y_2^m)$		
$Y_1 \setminus Y_2$	1	2	1	2	total
1	n_{111}	n_{112}	?	?	n_{21+}
2	n_{121}	n_{122}	?	?	n_{22+}

	$W=1\,(Y_1^o, Y_2^o)$		$W=2\,(Y_1^o, Y_2^m)$	
$Y_1 \setminus Y_2$	1	2	1	2
1	$1-\alpha_1$	$1-\alpha_2$	α_1	α_2
2	$1-\alpha_1$	$1-\alpha_2$	α_1	α_2

$$
\begin{aligned}
L(\boldsymbol{\theta}, \boldsymbol{\alpha} \,|\mathbf{N}) \propto\ & (\theta_{11}[1-\alpha_1])^{n_{111}} + (\theta_{12}[1-\alpha_2])^{n_{112}} + \\
& (\theta_{21}[1-\alpha_1])^{n_{121}} + ([1-\theta_{11}-\theta_{12}-\theta_{21}][1-\alpha_2])^{n_{122}} + \\
& (\theta_{11}\alpha_1 + \theta_{12}\alpha_2)^{n_{21+}} + (\theta_{21}\alpha_1 + [1-\theta_{11}-\theta_{12}-\theta_{21}]\alpha_2)^{n_{22+}}
\end{aligned}
$$

Estrut. MNAR sat. sem ajuste perfeito

	$W=1\,(Y_1^o, Y_2^o)$		$W=2\,(Y_1^o, Y_2^m)$		
$Y_1 \setminus Y_2$	1	2	1	2	total
1	n_{111}	n_{112}	?	?	n_{21+}
2	n_{121}	n_{122}	?	?	n_{22+}

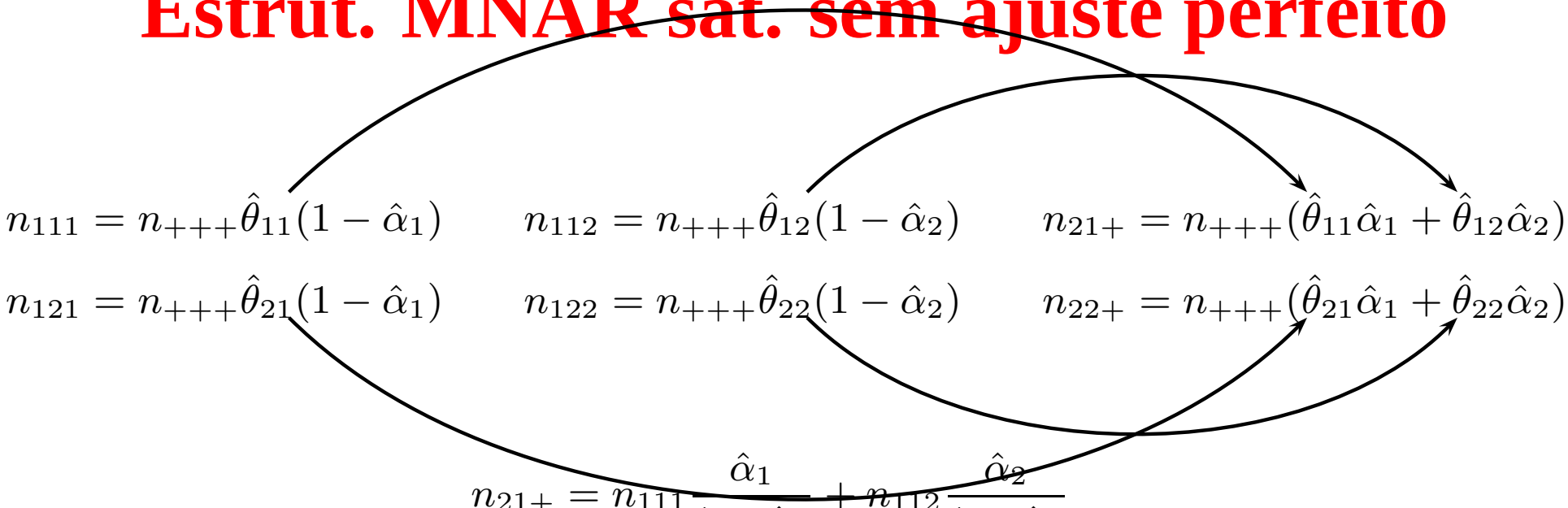
	$W=1\,(Y_1^o, Y_2^o)$		$W=2\,(Y_1^o, Y_2^m)$	
$Y_1 \setminus Y_2$	1	2	1	2
1	$1-\alpha_1$	$1-\alpha_2$	α_1	α_2
2	$1-\alpha_1$	$1-\alpha_2$	α_1	α_2

$$n_{111} = n_{+++}\hat{\theta}_{11}(1-\hat{\alpha}_1) \qquad n_{112} = n_{+++}\hat{\theta}_{12}(1-\hat{\alpha}_2) \qquad n_{21+} = n_{+++}(\hat{\theta}_{11}\hat{\alpha}_1 + \hat{\theta}_{12}\hat{\alpha}_2)$$

$$n_{121} = n_{+++}\hat{\theta}_{21}(1-\hat{\alpha}_1) \qquad n_{122} = n_{+++}\hat{\theta}_{22}(1-\hat{\alpha}_2) \qquad n_{22+} = n_{+++}(\hat{\theta}_{21}\hat{\alpha}_1 + \hat{\theta}_{22}\hat{\alpha}_2)$$

Baker e Laird (1988)

Estrut. MNAR sat. sem ajuste perfeito



$$n_{111} = n_{+++}\hat{\theta}_{11}(1-\hat{\alpha}_1) \qquad n_{112} = n_{+++}\hat{\theta}_{12}(1-\hat{\alpha}_2) \qquad n_{21+} = n_{+++}(\hat{\theta}_{11}\hat{\alpha}_1 + \hat{\theta}_{12}\hat{\alpha}_2)$$

$$n_{121} = n_{+++}\hat{\theta}_{21}(1-\hat{\alpha}_1) \qquad n_{122} = n_{+++}\hat{\theta}_{22}(1-\hat{\alpha}_2) \qquad n_{22+} = n_{+++}(\hat{\theta}_{21}\hat{\alpha}_1 + \hat{\theta}_{22}\hat{\alpha}_2)$$

$$n_{21+} = n_{111}\frac{\hat{\alpha}_1}{1-\hat{\alpha}_1} + n_{112}\frac{\hat{\alpha}_2}{1-\hat{\alpha}_2}$$

$$n_{22+} = n_{121}\frac{\hat{\alpha}_1}{1-\hat{\alpha}_1} + n_{122}\frac{\hat{\alpha}_2}{1-\hat{\alpha}_2}$$

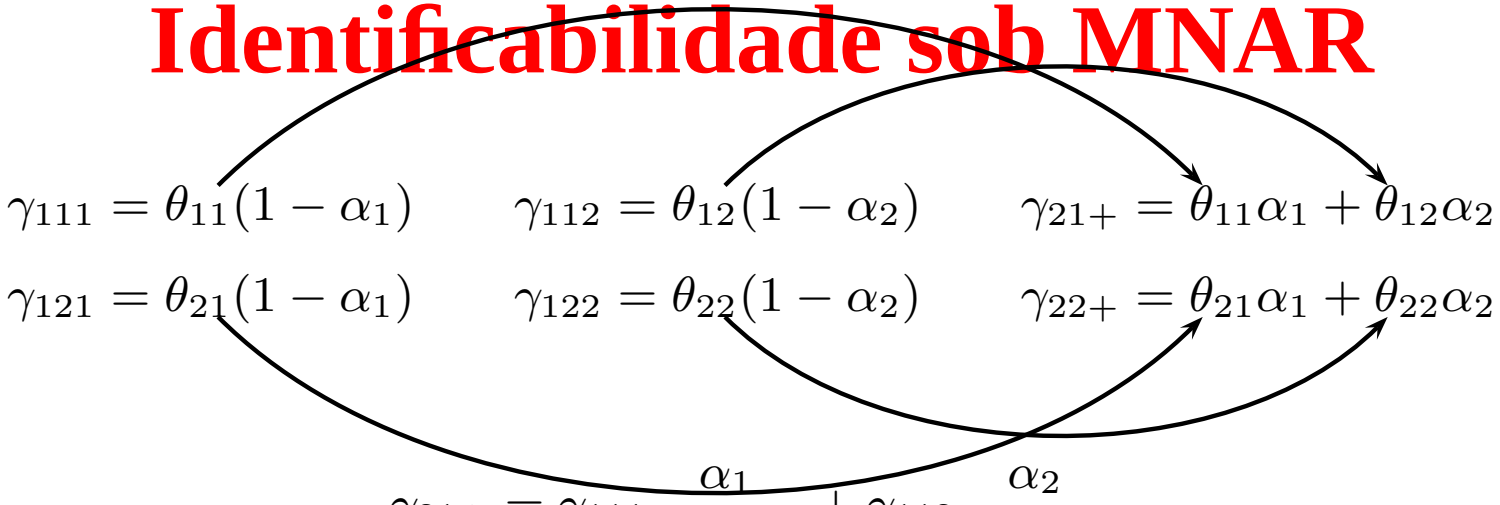
$$\hat{\alpha}_1 = 1 - \left(1 + \frac{n_{112}n_{22+} - n_{21+}n_{122}}{n_{112}n_{121} - n_{111}n_{122}}\right)^{-1} \qquad \hat{\alpha}_2 = 1 - \left(1 + \frac{n_{21+}n_{121} - n_{111}n_{22+}}{n_{112}n_{121} - n_{111}n_{122}}\right)^{-1}$$

$0 < \hat{\alpha}_1, \hat{\alpha}_2 < 1$ apenas se $\dfrac{n_{111}}{n_{121}} < \dfrac{n_{21+}}{n_{22+}} < \dfrac{n_{112}}{n_{122}}$

EMV sob MNAR

Caso	Condição	Nº máx.	Ajuste	$\alpha_1 = 0$ ou $\alpha_2 = 0$
I	$\frac{n_{111}}{n_{121}} < \frac{n_{21+}}{n_{22+}} < \frac{n_{112}}{n_{122}}$	1	perfeito	não
II	$\frac{n_{111}}{n_{121}} \neq \frac{n_{112}}{n_{122}}$ e $\frac{n_{111}}{n_{121}} = \frac{n_{21+}}{n_{22+}}$ ou $\frac{n_{112}}{n_{122}} = \frac{n_{21+}}{n_{22+}}$	1	perfeito	sim
III	$\frac{n_{111}}{n_{121}} < \frac{n_{112}}{n_{122}} < \frac{n_{21+}}{n_{22+}}$ ou $\frac{n_{21+}}{n_{22+}} < \frac{n_{111}}{n_{121}} < \frac{n_{112}}{n_{122}}$	1	imperfeito	sim
IV	$\frac{n_{111}}{n_{121}} = \frac{n_{112}}{n_{122}} \neq \frac{n_{21+}}{n_{22+}}$ e $n_{111} + n_{121} \neq n_{112} + n_{122}$	1	imperfeito	sim
V	$\frac{n_{111}}{n_{121}} = \frac{n_{112}}{n_{122}} = \frac{n_{21+}}{n_{22+}}$	∞	perfeito	não/sim
VI	$\frac{n_{111}}{n_{121}} = \frac{n_{112}}{n_{122}} \neq \frac{n_{21+}}{n_{22+}}$ e $n_{111} + n_{121} = n_{112} + n_{122}$	2	imperfeito	sim

Identificabilidade sob MNAR



$$\gamma_{111} = \theta_{11}(1-\alpha_1) \qquad \gamma_{112} = \theta_{12}(1-\alpha_2) \qquad \gamma_{21+} = \theta_{11}\alpha_1 + \theta_{12}\alpha_2$$

$$\gamma_{121} = \theta_{21}(1-\alpha_1) \qquad \gamma_{122} = \theta_{22}(1-\alpha_2) \qquad \gamma_{22+} = \theta_{21}\alpha_1 + \theta_{22}\alpha_2$$

$$\gamma_{21+} = \gamma_{111}\frac{\alpha_1}{1-\alpha_1} + \gamma_{112}\frac{\alpha_2}{1-\alpha_2}$$

$$\gamma_{22+} = \gamma_{121}\frac{\alpha_1}{1-\alpha_1} + \gamma_{122}\frac{\alpha_2}{1-\alpha_2}$$

$$\begin{pmatrix} \gamma_{21+} \\ \gamma_{22+} \end{pmatrix} = \begin{pmatrix} \gamma_{111} & \gamma_{112} \\ \gamma_{121} & \gamma_{122} \end{pmatrix} \begin{pmatrix} \frac{\alpha_1}{1-\alpha_1} \\ \frac{\alpha_2}{1-\alpha_2} \end{pmatrix}$$

para ter uma única solução para $\left(\frac{\alpha_1}{1-\alpha_1}, \frac{\alpha_2}{1-\alpha_2}\right)'$ deve satisfazer à condição

$$\frac{\gamma_{111}}{\gamma_{121}} \neq \frac{\gamma_{112}}{\gamma_{122}} \longleftrightarrow \frac{\theta_{11}\theta_{22}}{\theta_{21}\theta_{12}} \neq 1$$

Exemplo: tanto $(\theta_{11}, \theta_{12}, \theta_{21}, \alpha_1, \alpha_2)$ igual a $(0.1, 0.1, 0.4, 0.8, 0.3)$, como $(1/30, 1/6, 2/15, 0.4, 0.58)$, implicam

$$(\gamma_{111}, \gamma_{112}, \gamma_{121}, \gamma_{21+}, \gamma_{22+}) = (0.02, 0.07, 0.08, 0.11, 0.44)$$

Estudo de simulação

100 mil réplicas de Monte Carlo

$n_{+++} = 50,\ 100,\ 200,\ 500,\ 1\,000,\ 2\,000,\ 5\,000,\ 10\,000,\ 20\,000,\ 50\,000,\ 100\,000$

$(\alpha_1, \alpha_2) = (0.8, 0.3)$

Estruturas para $\{\theta_{ij}\}$

	A1		B1		C1		D1	
$Y_1 \setminus Y_2$	1	2	1	2	1	2	1	2
1	0.30	0.20	0.20	0.10	0.30	0.20	0.30	0.10
2	0.20	0.30	0.20	0.50	0.40	0.10	0.40	0.20

	A2		B2		C2		D2	
$Y_1 \setminus Y_2$	1	2	1	2	1	2	1	2
1	0.25	0.25	0.20	0.20	0.15	0.15	0.10	0.10
2	0.25	0.25	0.30	0.30	0.35	0.35	0.40	0.40

Estimativas de P(ajuste perfeito) e $E(\hat{\theta}_{21} - \hat{\theta}_{12})$

Proporções de réplicas (%) com

EMV dentro do espaço paramétrico e ajuste perfeito

n_{+++}	Estruturas							
	A1	B1	C1	D1	A2	B2	C2	D2
50	50.1	62.4	50.5	43.9	41.2	41.3	41.5	42.1
100	57.3	71.6	56.5	46.0	41.7	41.8	41.9	41.9
200	65.2	80.9	62.0	48.9	41.8	42.0	41.8	41.7
500	76.6	91.9	69.4	54.7	42.1	41.9	42.0	42.3
1 000	84.4	97.7	76.4	60.5	42.2	42.2	42.0	42.2
2 000	92.6	99.8	84.2	65.6	42.4	42.3	42.2	42.0
5 000	98.9	100.0	94.5	73.7	42.1	42.0	42.3	42.2
10 000	99.9	100.0	98.8	81.6	42.2	42.3	42.2	42.1
20 000	100.0	100.0	99.9	89.9	42.0	42.0	42.2	42.4
50 000	100.0	100.0	100.0	97.8	42.3	41.8	42.2	42.2
100 000	100.0	100.0	100.0	99.8	42.0	42.2	42.7	42.2

Estimativas de Monte Carlo dos valores esperados dos EMV de $\theta_{21} - \theta_{12}$ quando as EMV estão dentro do espaço paramétrico e ajuste perfeito

n_{+++}	Estruturas							
	A1	B1	C1	D1	A2	B2	C2	D2
50	−0.149	0.012	−0.002	0.050	−0.182	−0.079	0.018	0.118
100	−0.122	0.043	0.034	0.063	−0.176	−0.075	0.025	0.125
200	−0.091	0.073	0.079	0.085	−0.172	−0.071	0.027	0.128
500	−0.047	0.097	0.133	0.125	−0.170	−0.071	0.029	0.129
1 000	−0.021	0.104	0.162	0.167	−0.170	−0.070	0.031	0.130
2 000	−0.005	0.104	0.182	0.210	−0.171	−0.071	0.030	0.130
5 000	0.003	0.102	0.196	0.255	−0.172	−0.070	0.030	0.130
10 000	0.002	0.101	0.200	0.277	−0.169	−0.072	0.031	0.130
20 000	0.001	0.100	0.201	0.291	−0.171	−0.069	0.030	0.129
50 000	0.000	0.100	0.200	0.299	−0.171	−0.069	0.030	0.130
100 000	0.000	0.100	0.200	0.301	−0.170	−0.070	0.030	0.131
$\theta_{21} - \theta_{12}$	0.000	0.100	0.200	0.300	0.000	0.100	0.200	0.300
ACC	−0.222	−0.060	−0.171	0.029	−0.278	−0.178	−0.078	0.022

ACC: $P(Y_1 = 2, Y_2 = 1|W = 1) - P(Y_1 = 1, Y_2 = 2|W = 1)$

Seleção do mecanismo de omissão

- Murray e Findlay (1988) descreveram um estudo de hipertensão que a adoção do mecanismo MAR se justifica pelo planejamento experimental
- Baseia-se em suposições inverificáveis

	$W=1\ (Y_1^o, Y_2^o)$		$W=2\ (Y_1^o, Y_2^m)$			$W=3\ (Y_1^m, Y_2^o)$		$W=4\ (Y_1^m, Y_2^m)$		
$Y_1 \setminus Y_2$	1	2	1	2	total	1	2	1	2	total
1	n_{111}	n_{112}	?	?	n_{21+}	?	?	?	?	
2	n_{121}	n_{122}	?	?	n_{22+}	?	?	?	?	
total						n_{3+1}	n_{3+2}			n_{4++}

- Acompanhar o estudo e investigar o motivo da omissão pode sugerir estruturas “mais adequadas”
- Análise de sensibilidade:
  - da estabilidade das inferências de interesse e
  - da plausibilidade dos valores esperados estimados para as freqüências ampliadas $\widehat{E(n_{tij})} = n_{+++}\hat{\theta}_{ij}\lambda_{t(ij)}(\hat{\boldsymbol{\alpha}})$

Análise de sensibilidade

- Molenberghs, Kenward e Goetghebeur (2001) distinguiram 2 tipos de **incertezas estatísticas**:
  - **imprecisão estatística**, devido à amostragem
  - **ignorância estatística**, causada pela omissão
- Conforme $n_{+++} \longrightarrow \infty$, a imprecisão desaparece e resta a ignorância
- A imprecisão estatística pode ser capturada por erros padrões e regiões de confiança
- A ignorância estatística deve ser avaliada por **regiões de ignorância** e **regiões de incerteza**
- Obtidas por meio de modelos sobre-parametrizados

Passos da análise de sensibilidade

- Propõe-se um modelo sobre-parametrizado
- Particiona-se os parâmetros em $(\boldsymbol{\mu}, \boldsymbol{\tau})$, de tal forma que
  - a dimensão de $\boldsymbol{\mu}$ (**parâmetro estimável**) seja o número de graus de liberdade dos dados observados
  - $\boldsymbol{\tau}$ (**parâmetro de sensibilidade**) tenha um ou mais parâmetros selecionados dentre os restantes
- Cada valor fixado de $\boldsymbol{\tau}$ produz uma estimativa $\hat{\boldsymbol{\mu}}(\boldsymbol{\tau})$ de $\boldsymbol{\mu}$ e uma região de $100(1-\alpha)\%$ confiança
- A união das estimativas pontuais gera a região de **ignorância** para $\boldsymbol{\mu}$
- A união das regiões de confiança produz a região de $100(1-\alpha)\%$ **incerteza** para $\boldsymbol{\mu}$

Intervalo para o melhor-pior caso

(*best-worst case interval*)

- Alocar as unidades omissas em categorias que produzam casos extremos para as inferências de interesse
- Possui grande amplitude
- Método simples, bastante informativo e um ponto de partida honesto para uma modelagem cautelosa

Ex.de intervalo para o melhor-pior caso

Objetivo: avaliar a associação entre o hábito de fumo da mãe e o peso do recém-nascido (razão de chances)

Dados observados

MFum \ PesoRN	$W=1$		$W=2$			$W=3$		$W=4$			Total		
	<2.5	≥ 2.5	<2.5	≥ 2.5	total	<2.5	≥ 2.5	<2.5	≥ 2.5	total	<2.5	≥ 2.5	total
sim	4512	21009	?	?	1049	?	?	?	?		?	?	
não	3394	24132	?	?	1135	?	?	?	?		?	?	
total						142	464			1224			57061

Dados observados

MFum \ PesoRN	$W=1$		$W=2$			$W=3$		$W=4$			Total		
	<2.5	≥ 2.5	<2.5	≥ 2.5	total	<2.5	≥ 2.5	<2.5	≥ 2.5	total	<2.5	≥ 2.5	total
sim	4512	21009	?	?	1049	?	?	?	?		?	?	
não	3394	24132	?	?	1135	?	?	?	?		?	?	
total						142	464			1224			57061

Caso extremo com a menor estimativa possível para a razão de chances: 0.82

sim	4512	21009	0	1049	1049	0	464	0	0		4512	22522	
não	3394	24132	1135	0	1135	142	0	1224	0		5895	24132	
total						142	464			1224			57061

Dados observados

Estruturas MNAR para o exemplo

(Y_1: hábito de fumo da mãe e Y_2: do peso do recém-nascido)

- $\psi_{1(ij)} = P(Y_1^o | Y_1 = i, Y_2 = j)$
  $\psi_{21(ij)} = P(Y_2^o | Y_1^o, Y_1 = i, Y_2 = j)$
  $\psi_{20(ij)} = P(Y_2^o | Y_1^m, Y_1 = i, Y_2 = j)$

- $\lambda_{1(ij)} = \psi_{1(ij)}\psi_{21(ij)}$
  $\lambda_{2(ij)} = \psi_{1(ij)}\left(1 - \psi_{21(ij)}\right)$
  $\lambda_{3(ij)} = \left(1 - \psi_{1(ij)}\right)\psi_{20(ij)}$
  $\lambda_{4(ij)} = \left(1 - \psi_{1(ij)}\right)\left(1 - \psi_{20(ij)}\right)$

- $\text{logito}\left(\psi_{1(ij)}\right) = \alpha_{10} + \alpha_1(i-1) + \alpha_2(j-1)$ (MNAR1)
  $\text{logito}\left(\psi_{21(ij)}\right) = \alpha_{20} + \alpha_1(i-1) + \alpha_2(j-1)$
  $\text{logito}\left(\psi_{20(ij)}\right) = \alpha_{30} + \alpha_1(i-1) + \alpha_2(j-1)$

- $\text{logito}\left(\psi_{1(ij)}\right) = \alpha_{10} + \alpha_1(i-1) + \alpha_2(j-1) + \alpha_3(i-1)(j-1)$ (MNAR2)
  $\text{logito}\left(\psi_{21(ij)}\right) = \alpha_{20} + \alpha_1(i-1) + \alpha_2(j-1) + \alpha_3(i-1)(j-1)$
  $\text{logito}\left(\psi_{20(ij)}\right) = \alpha_{20} + \alpha_1(i-1) + \alpha_2(j-1) + \alpha_3(i-1)(j-1)$

Resultados das análises

EMV e intervalos de 95% de confiança para a razão de chances

ACC	1.53	(1.46; 1.60)
MCAR / MAR	1.53	(1.46; 1.60)
MNAR1	1.50	(1.42; 1.57)
MNAR2	0.83	(0.79; 0.86)

MCAR

MFum \ PesoRN	$W=1$		$W=2$			$W=3$		$W=4$			Total		
	<2.5	≥ 2.5	<2.5	≥ 2.5	total	<2.5	≥ 2.5	<2.5	≥ 2.5	total	<2.5	≥ 2.5	total
sim	4541	20984	187	864	1051	52	240	105	484		4884	22571	27456
não	3416	24106	141	992	1133	39	275	79	556		3675	25930	29605
total						91	515			1224	8559	48502	57061

MAR

sim	4512	21009	187	862	1049	81	216	105	484		4884	22571	27456
não	3394	24132	141	994	1135	61	248	79	556		3675	25930	29605
total						142	464			1224	8559	48502	57061

MNAR1

sim	4512	21009	525	524	1049	80	208	379	212		5496	21954	27449
não	3394	24132	450	685	1135	62	256	336	296		4242	25369	29612
total						142	464			1224	9738	47323	57061

MNAR2

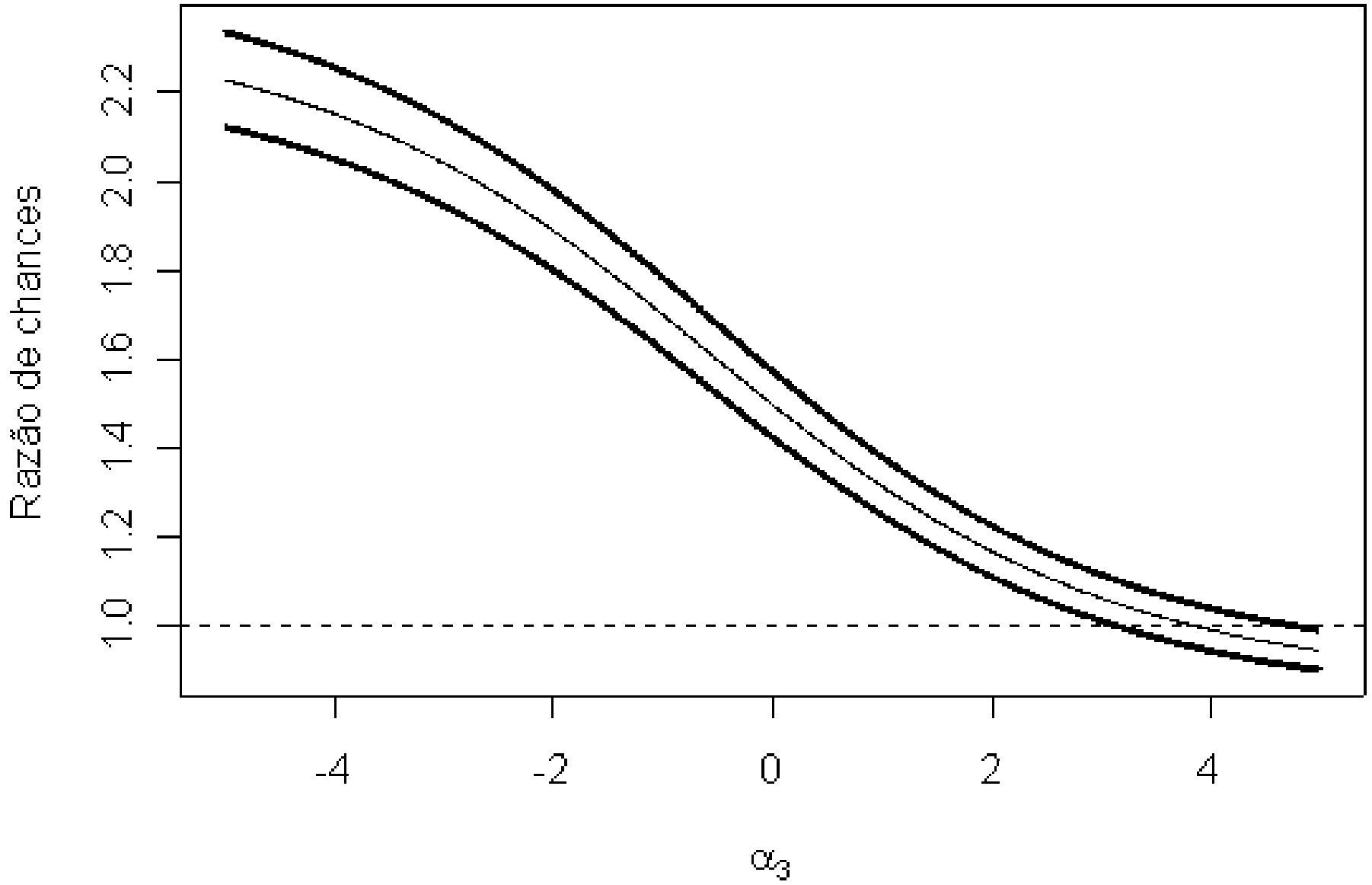
sim	4512	20977	0	1088	1088	0	489	0	25		4512	22579	27091
não	2744	24132	1778	0	1778	799	0	517	0		5838	24132	29970
total						799	489			543	10350	46711	57061

MFum: Mãe fumante, PesoRN: Peso do recém-nascido (kg)

Estrutura MNAR sobre-parametrizada

- $$\begin{aligned}\text{logito}\left(\psi_{1(ij)}\right) &= \alpha_{10} + \alpha_1(i-1) + \alpha_2(j-1) + \alpha_3(i-1)(j-1)\\ \text{logito}\left(\psi_{21(ij)}\right) &= \alpha_{20} + \alpha_1(i-1) + \alpha_2(j-1) + \alpha_3(i-1)(j-1)\\ \text{logito}\left(\psi_{20(ij)}\right) &= \alpha_{30} + \alpha_1(i-1) + \alpha_2(j-1) + \alpha_3(i-1)(j-1)\end{aligned}$$
- α_3 é o parâmetro de sensibilidade e os demais, estimáveis
- Quanto maior é α_3, maiores são as chances de se observar Y_1 e Y_2 quando tiverem conjuntamente o valor 2
- Logo, menos unidades com omissão são alocadas em $(Y_1 = 2, Y_2 = 2)$ e menores são as razões de chance obtidas
- Intervalo de ignorância para a razão de chances é (0.94; 2.23)
- Intervalo de 95% de incerteza para a razão de chances é (0.90; 2.34)

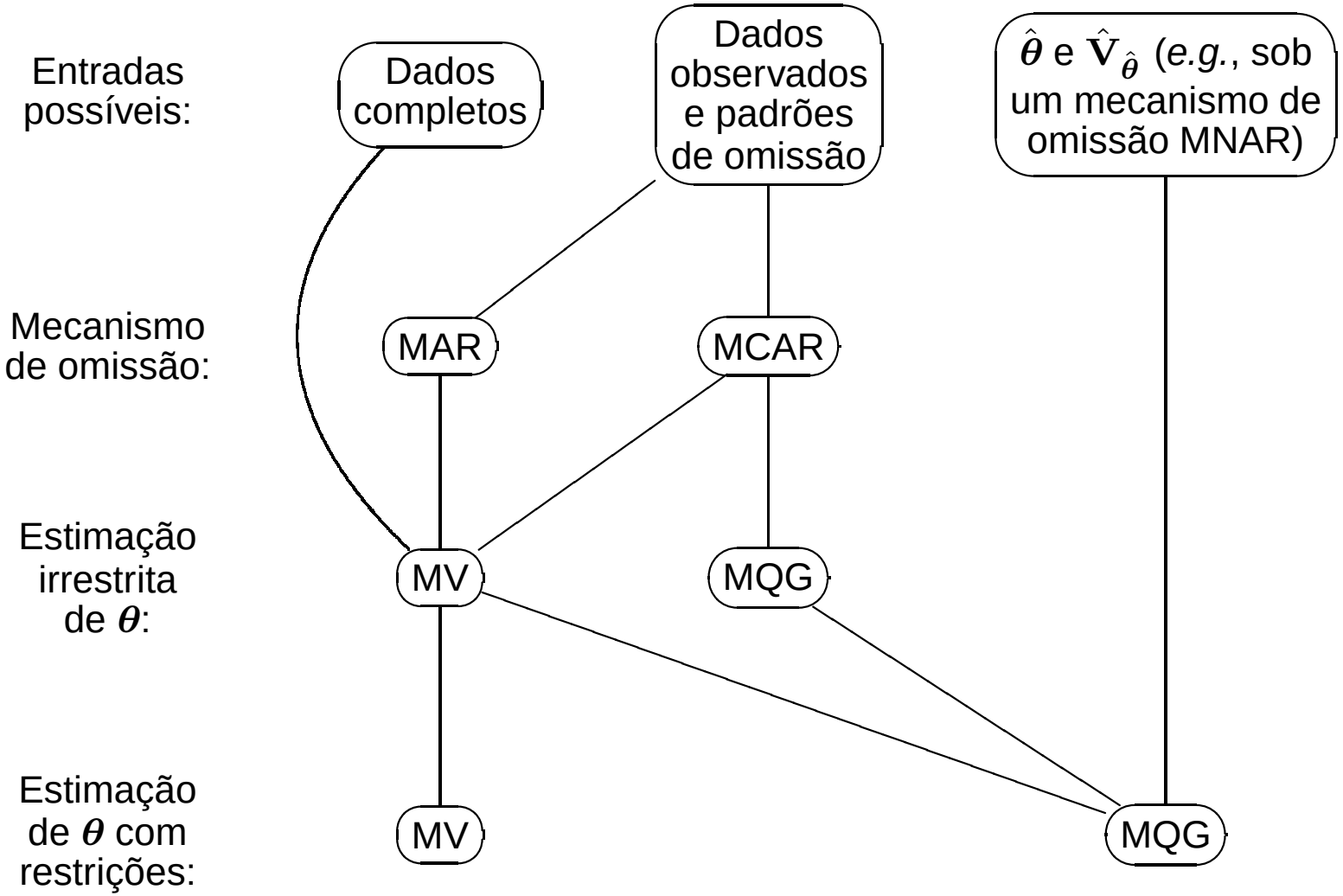
Análise de sensibilidade



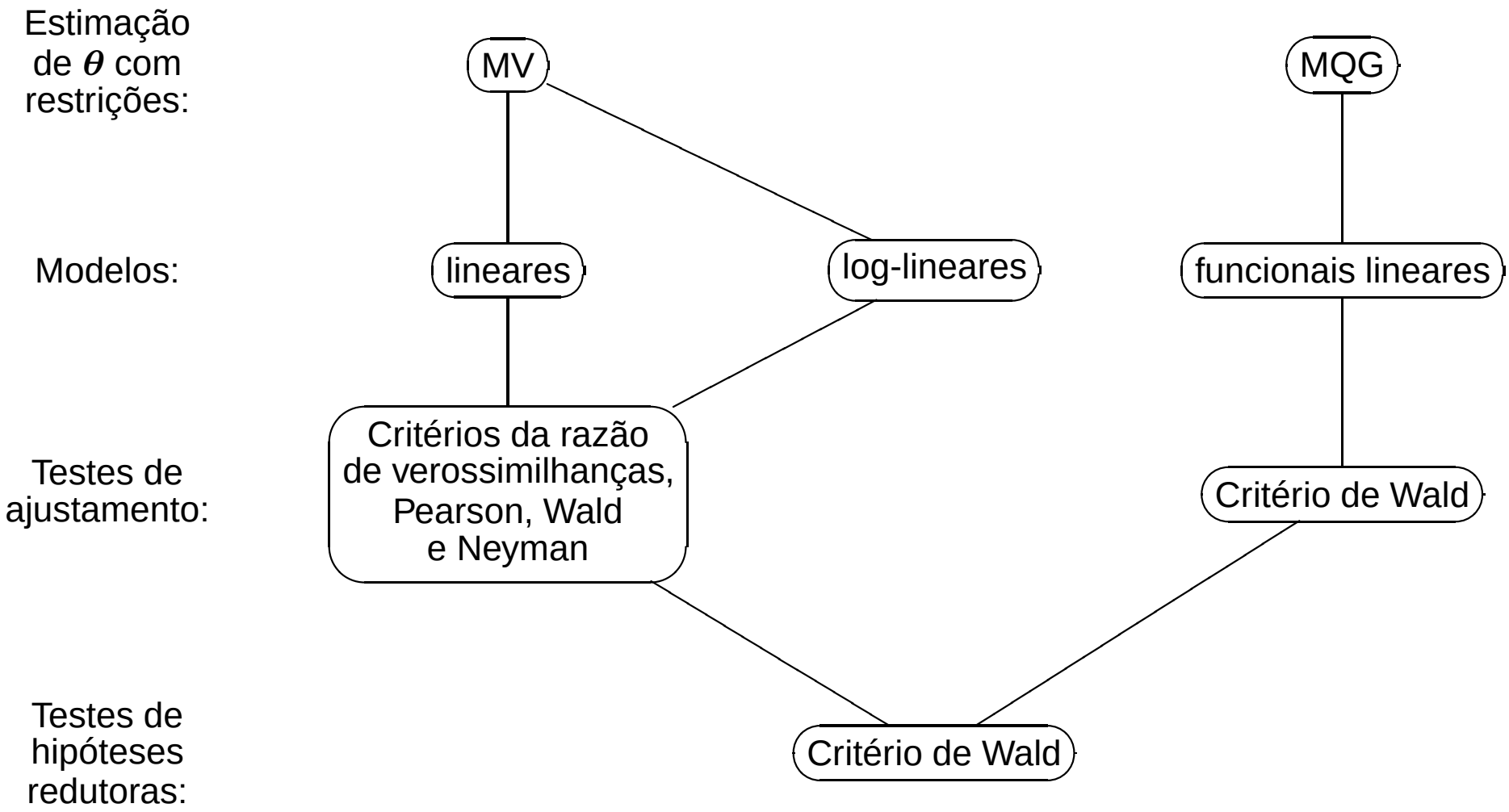
Procedimentos inferenciais

- Paulino (1988, 1991) e Paulino e Singer (2006) apresentam fórmulas matriciais para o ajuste de modelos lineares e log-lineares multinomiais por MV sob MAR e MCAR e funcionais lineares por MQG sob MCAR
- Também propõem uma metodologia híbrida que consiste em
  - primeiramente ajustar um modelo saturado para as probabilidades de categorização sob MAR ou MCAR por MV e usar suas EMV e matriz de covariâncias num
  - segundo estágio para ajustar modelos por MQG, no espírito da regressão funcional assintótica
- Poleto (2006) e Poleto, Singer e Paulino (2007a) estendem a modelagem para o caso em que há subpopulações (produto de multinomiais)

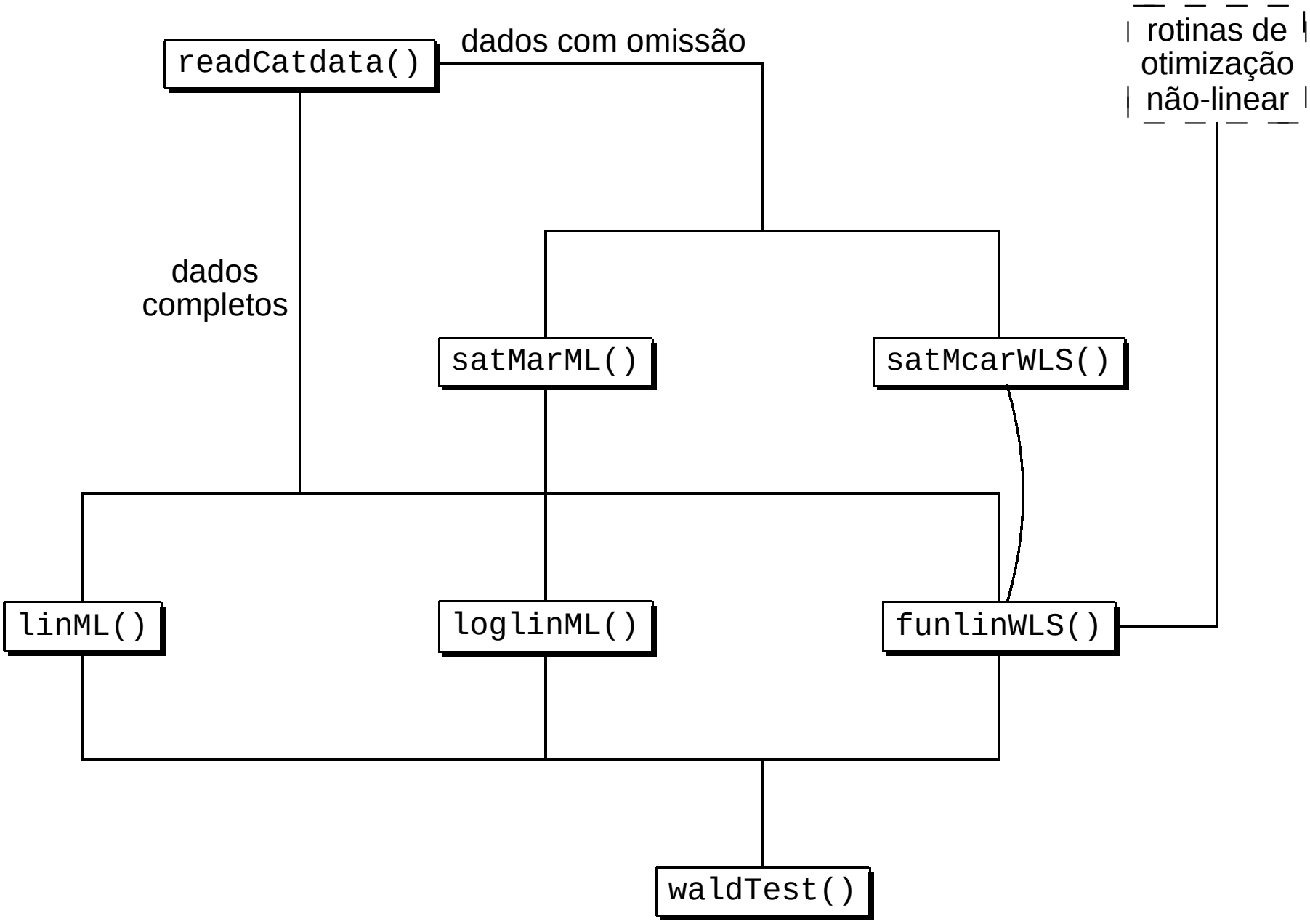
Análises realizadas com a biblioteca `Catdata`



Análises realizadas com a biblioteca `Catdata`



Biblioteca `Catdata` de rotinas para o R



Outros exemplos de ADC com omissão

Poleto (2006)

- Ajuste de modelos log-lineares (razão de chances adjacentes comum) com 2 subpopulações, ambas com omissão
- Padrão de omissão resultante do confundimento de células vizinhas
  Ajuste de modelos lineares (homogeneidade marginal) e funcionais lineares (*kappa*, *kappa* ponderado)
- Estudo de viés de não-resposta em pesquisas eleitorais
- Ajuste de modelos funcionais lineares (sensibilidade/especificidade, valor preditivo positivo/negativo) para a comparação da precisão de testes diagnósticos (também em Poleto, Singer e Paulino, 2007b)
  Sem evidências contra o mecanismo MCAR, mas conclusões sob a ACC diferentes, devido à grande quantidade de omissões

Extensão da notação para dados com omissão

- $\{n_{stc}\}$ indicam as unidades da s-ésima subpopulação com o t-ésimo padrão de omissão classificadas na c-ésima classe de resposta, $s = 1, \ldots, S$, $t = 1, \ldots, T_s$, $c = 1, \ldots, R_{st}$

- O cenário sem omissão ($t = 1$) possui classes equivalentes às R categorias de respostas, *i.e.*, $R_{s1} = R_1 = R$

- T_s é o nº de cenários de omissão da s-ésima subpopulação

- Criação de vetores $\mathbf{z}_{stc}$ (indicadores de respostas), de dimensão $R \times 1$, com elementos iguais a 1 para os componentes correspondentes aos parâmetros do vetor $\boldsymbol{\theta}_s = \left(\theta_{1(s)}, \ldots, \theta_{R(s)}\right)'$ que estiverem associados às freqüências n_{stc} e com demais elementos nulos

- $\mathbf{Z}_{st} = (\mathbf{z}_{st1}, \ldots, \mathbf{z}_{stR_{st}})$ é uma matriz $R \times R_{st}$ englobando os vetores indicadores para o t-ésimo padrão de omissão da s-ésima subpopulação

Especificação dos padrões de omissão

- O argumento Zp deve receber

$$\left[(\mathbf{Z}_{s2}, \ldots, \mathbf{Z}_{sT_s}), s = 1, \ldots, S\right]$$

- Recupera-se cada uma das submatrizes $\mathbf{Z}_{st}$, utilizando o argumento Rp, que deve conter em cada uma de suas $s = 1, \ldots, S$ linhas

$$R_{s2}, \ldots, R_{sT_s}$$

- TF deve receber em cada uma das $s = 1, \ldots, S$ linhas

$$n_{s11}, \ldots, n_{s1R}, n_{s21}, \ldots, n_{s2R_2}, \ldots, n_{sT_s1}, \ldots, n_{sT_sR_{T_s}}$$

Problema da avaliação da função pulmonar

- Estudo da FM-USP com 167 crianças e adolescentes asmáticos
- Aos 5 e 7 minutos de um teste cicloergométrico, verificou-se se o indivíduo apresentou broncoespasmo induzido pelo exercício (BIE)
- Objetivo: avaliar se a distribuição do BIE varia com o tempo

BIE	BIE aos 7 min.		
aos 5 min.	sim	não	omisso
sim	12	4	50
não	5	2	31
omisso	27	12	24

Paulino e Singer (2006), Exemplo 13.4

Problema da avaliação da função pulmonar

BIE aos 5 min.	BIE aos 7 min.		
	sim	não	omisso
sim	12	4	50
não	5	2	31
omisso	27	12	24

BIE aos 5 min.	BIE aos 7 min.	
	sim	não
sim	θ_{11}	θ_{12}
não	θ_{21}	θ_{22}

BIE aos 5 min.	BIE aos 7 min.		
	$j=1$	$j=2$	omisso
$i=1$ $i=2$	$t=1$		$t=2$
omisso	$t=3$		$t=4$

Não há subpop., $\therefore$ elimina-se s

Zp=$(\mathbf{Z}_2, \mathbf{Z}_3) = (\mathbf{z}_{21}, \mathbf{z}_{22}, \mathbf{z}_{31}, \mathbf{z}_{32})$

$$= \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \qquad \boldsymbol{\theta} = \begin{pmatrix} \theta_{11} \\ \theta_{12} \\ \theta_{21} \\ \theta_{22} \end{pmatrix}$$

$$50 \quad 31 \quad 27 \quad 12$$

Rp=$(R_2, R_3) = (2, 2)$

TF=$(12, 4, 5, 2, 50, 31, 27, 12)$

Cenários de omissão completa ($t = 4$) não trazem informação para a estimação de $\boldsymbol{\theta}$ sob MAR e MCAR, então a freqüência $n_{41} = 24$ é ignorada

Função de verossimilhança do MNAR1

$$
\begin{aligned}
L\left(\boldsymbol{\theta}, \boldsymbol{\alpha} \,|\mathbf{N}\right) \propto & \prod_{i=1}^{2}\prod_{j=1}^{2}\left(\theta_{ij}\lambda_{1(ij)}\right)^{n_{1ij}} \times \prod_{i=1}^{2}\left(\theta_{i1}\lambda_{2(i1)} + \theta_{i2}\lambda_{2(i2)}\right)^{n_{2i+}} \times \\
& \prod_{j=1}^{2}\left(\theta_{1j}\lambda_{3(1j)} + \theta_{2j}\lambda_{3(2j)}\right)^{n_{3+j}} \times \left(\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}\lambda_{4(ij)}\right)^{n_{4++}} \\
= & \prod_{i=1}^{2}\prod_{j=1}^{2}\left(\theta_{ij}\psi_{1(ij)}\psi_{21(ij)}\right)^{n_{1ij}} \times \prod_{i=1}^{2}\left(\sum_{j=1}^{2}\theta_{ij}\psi_{1(ij)}\left(1-\psi_{21(ij)}\right)\right)^{n_{2i+}} \times \\
& \prod_{j=1}^{2}\left(\sum_{i=1}^{2}\theta_{ij}\left(1-\psi_{1(ij)}\right)\psi_{20(ij)}\right)^{n_{3+j}} \times \\
& \left(\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}\left(1-\psi_{1(ij)}\right)\left(1-\psi_{20(ij)}\right)\right)^{n_{4++}}
\end{aligned}
$$

Função de verossimilhança do MNAR1

$$L(\boldsymbol{\theta}, \boldsymbol{\alpha} \,|\mathbf{N}) \propto \prod_{i=1}^{2}\prod_{j=1}^{2}\left(\theta_{ij}\psi_{1(ij)}\psi_{21(ij)}\right)^{n_{1ij}} \times \prod_{i=1}^{2}\left(\sum_{j=1}^{2}\theta_{ij}\psi_{1(ij)}\left(1-\psi_{21(ij)}\right)\right)^{n_{2i+}} \times$$

$$\prod_{j=1}^{2}\left(\sum_{i=1}^{2}\theta_{ij}\left(1-\psi_{1(ij)}\right)\psi_{20(ij)}\right)^{n_{3+j}} \times$$

$$\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}\left(1-\psi_{1(ij)}\right)\left(1-\psi_{20(ij)}\right)\right)^{n_{4++}}$$

$$= \prod_{i=1}^{2}\prod_{j=1}^{2}\left(\theta_{ij}\frac{\mathrm{e}^{\alpha_{10}+\alpha_1(i-1)+\alpha_2(j-1)}}{1+\mathrm{e}^{\alpha_{10}+\alpha_1(i-1)+\alpha_2(j-1)}}\frac{\mathrm{e}^{\alpha_{20}+\alpha_1(i-1)+\alpha_2(j-1)}}{1+\mathrm{e}^{\alpha_{20}+\alpha_1(i-1)+\alpha_2(j-1)}}\right)^{n_{1ij}} \times$$

$$\prod_{i=1}^{2}\left(\sum_{j=1}^{2}\theta_{ij}\frac{\mathrm{e}^{\alpha_{10}+\alpha_1(i-1)+\alpha_2(j-1)}}{1+\mathrm{e}^{\alpha_{10}+\alpha_1(i-1)+\alpha_2(j-1)}}\frac{1}{1+\mathrm{e}^{\alpha_{20}+\alpha_1(i-1)+\alpha_2(j-1)}}\right)^{n_{2i+}} \times$$

$$\prod_{j=1}^{2}\left(\sum_{i=1}^{2}\theta_{ij}\frac{1}{1+\mathrm{e}^{\alpha_{10}+\alpha_1(i-1)+\alpha_2(j-1)}}\frac{\mathrm{e}^{\alpha_{30}+\alpha_1(i-1)+\alpha_2(j-1)}}{1+\mathrm{e}^{\alpha_{30}+\alpha_1(i-1)+\alpha_2(j-1)}}\right)^{n_{3+j}} \times$$

$$\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\theta_{ij}\frac{1}{1+\mathrm{e}^{\alpha_{10}+\alpha_1(i-1)+\alpha_2(j-1)}}\frac{1}{1+\mathrm{e}^{\alpha_{30}+\alpha_1(i-1)+\alpha_2(j-1)}}\right)^{n_{4++}}$$