Análise de Dados Longitudinais
Modelos Lineares Generalizados Longitudinais

Enrico A. Colosimo/UFMG

Respostas Longitudinal Não-Gaussiana

1. $Y_{ij}, \quad i = 1, \ldots, N; \; j = 1, \ldots, k$: binária, contagem, etc.
2. Modelos Estatísticos
   - Modelos Lineares Generalizados Mistos.
   - Modelos Marginais: GEE

Exemplos

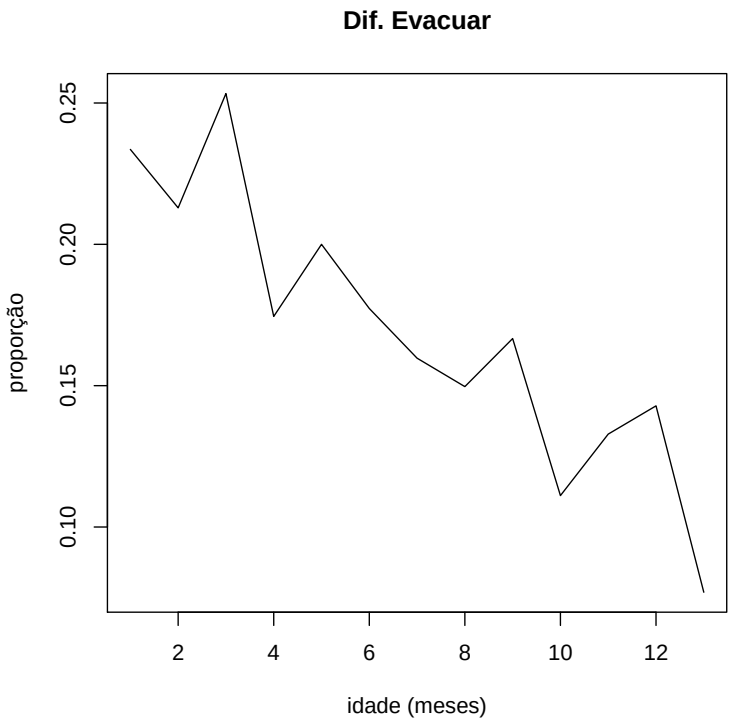
1. Mecanismo Evacuatório de Récem-Nascidos
2. Fatores de Risco Coronariano: MCRF, (FLW, pag. 364)

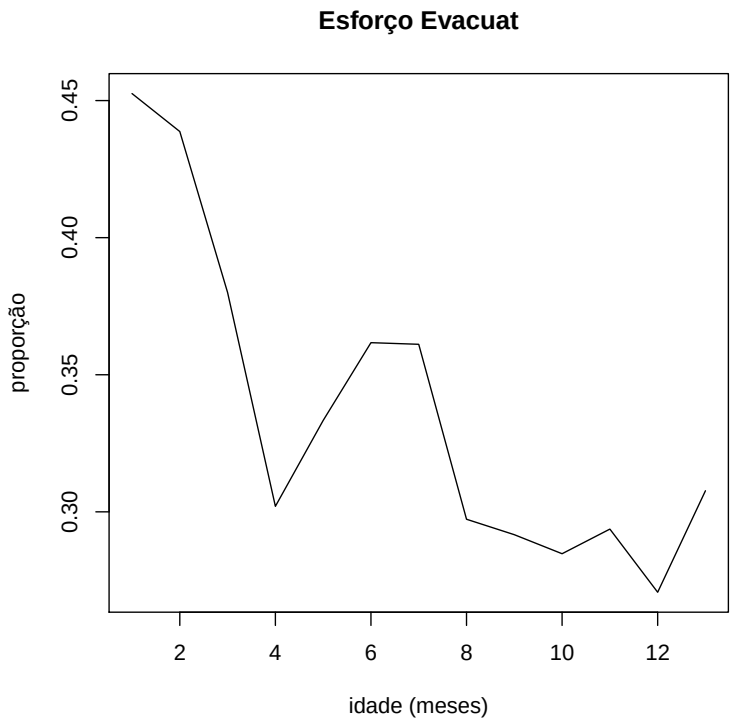
Mecanismo Evacuatório de Récem-Nascidos

- 151 recém-nascidos acompanhados nos primeiros 12 meses de vida no Hospital das Clínicas da UFMG em 2010 e 2011.
- Acompanhamento mensal totalizando 1751 medidas (61 perdas)
- Respostas: (1) Binárias: Dificuldade para evacuar, Esforço evacuatório, Dor ao evacuar e (2) Contagem: Freqüência evacuatória/semana.
- Variável temporal: idade (em dias ou meses).
- Covariáveis: 1- fixa (sexo) e 2- dependentes do tempo: Aleitamento materno, dieta (0/1): cereais; frutas; vegetais, carnes, etc.
- Objetivo: avaliar o comportamento temporal das respostas e seus respectivos indicadores.

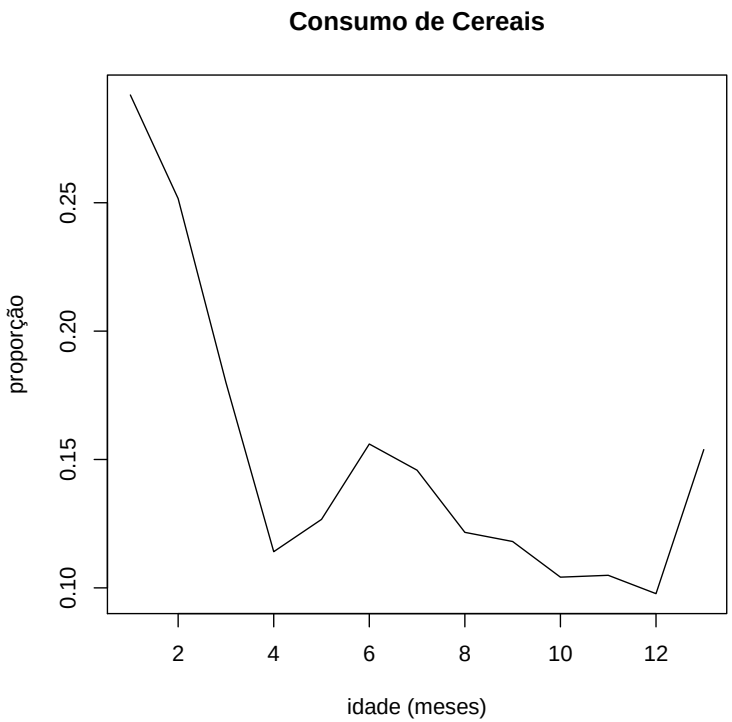
Resposta: Dificuldade e Esforço para Evacuar

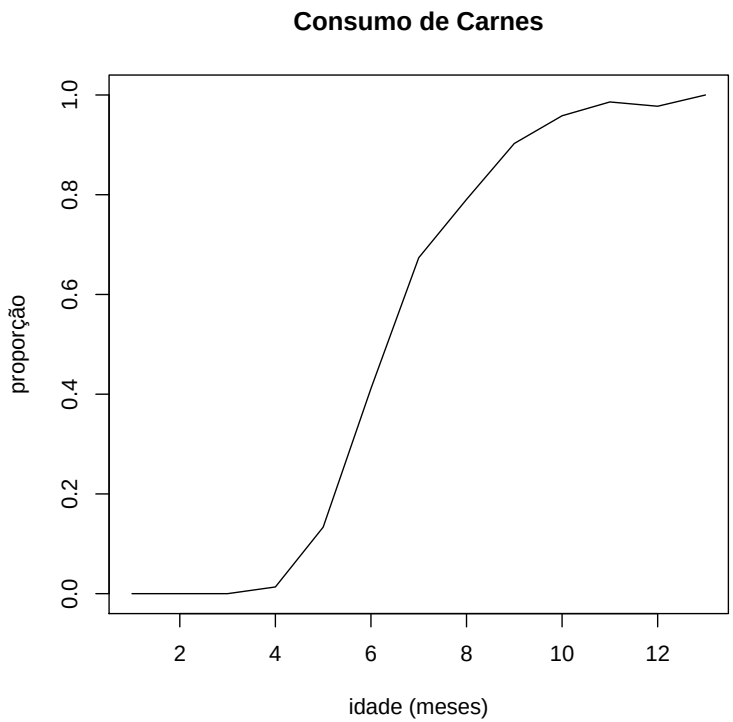
Obs.: idade foi arrendonda para mês (um único digito).





Covariáveis: Consumo de Cereais e de Carnes





”Muscatine Coronary Risk Factor Study”

- Estudo longitudinal de crianças em idade escolar realizado em Muscatine, Iowa, Estados Unidos na década de 80.
- Cinco coortes de crianças, inicialmente com idades em 5-7, 7-9, 9-11, 11-13 e 13-15 foram acompanhadas bianualmente de 1977 a 1981 (3 medidas).
- Respostas binária: obesidade.
- Variável temporal: idade (em dias ou meses).
- Covariável: sexo.
- Objetivo: avaliar (1) se o risco de obesidade aumenta com a idade e (2) se os padrões são os mesmos para meninos e meninas.

”Muscatine Coronary Risk Factor Study”

Gênero	Coorte Idade	Obesidade (%) 1977	1979	1981
Meninos				
	5-7	7.9	15.4	21.2
	7-9	18.8	20.5	23.7
	9-11	21.2	22.7	22.5
	11-13	24.3	21.8	19.4
	13-15	19.2	21.1	18.2
Meninas				
	5-7	14.0	17.2	25.1
	7-9	16.5	24.0	24.9
	9-11	25.4	26.2	22.2
	11-13	23.8	22.1	19.9
	13-15	22.9	25.8	20.9

Revisão: Modelos Lineares Generalizados

Modelos Lineares Generalizados (MLG) é uma classe unificada de modelos de Regressão.

1. Considere $Y_1, \ldots Y_N$ uma amostra aleatória de respostas univariadas (desenho transversal).

2. Um vetor de p-covariáveis associados a cada resposta Y_i. Ou seja

$$X_i = \begin{pmatrix} X_{i0} \\ X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix}$$

em que $X_{i0} = 1$.

3 O MLG é definido por três componentes:

- Distribuição de Y_i.
- Componente Sistemático (preditor linear).

$$\eta_i = X_i'\beta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Função de Ligação.

MLG - Família Exponencial

A distribuição de Y_i pertence a família exponencial que inclue os principais modelos estatísticos: normal, binomial, poisson, exponencial, etc.

Ou seja,
Y_i tem densidade $f(Y_i|\theta, \phi)$ pertencente a família exponencial.

$$f(y_i|\theta_i, \phi) = \exp\{\phi^{-1}(y_i\theta_i - \psi(\theta_i)) + c(y_i, \phi)\}$$

em que θ_i é parâmetro natural, ϕ é o de escala e específicas funções $\psi(.)$ e $c(.)$.

- $\psi(.)$ é a função geradora de momentos
  - $\mu = E(Y) = \psi'(\theta)$ e
  - $Var(Y) = \phi\psi''(\theta)$
- Em geral, média e variância são relacionadas.

$$Var(Y) = \phi\psi'' \quad (\psi'^{-1}(\mu) = \phi\nu(\mu))$$

- A função $\nu(\mu)$ é chamada de função de variância.
- ψ'^{-1} que relaciona θ com μ é chamada de função de ligação.

Exemplos

1. Modelo Normal (μ, σ^2)
   - $\theta = \mu$
   - $\phi = \sigma^2$
   - $\psi(\theta) = \theta^2/2$
   - Média: $\mu = \theta$ e $\nu(\mu) = 1$
   - Observe que no modelo normal, média e variância não são relacionadas
     $$\phi\nu(\mu) = \sigma^2$$
   - Função de ligação natural: $\theta = \mu$.

Exemplos

2 Modelo Bernoulli (π)

- $\theta = \log(\pi/(1-\pi)$
- $\phi = 1$
- $\psi(\theta) = \log(1-\pi) = log(1+\exp(\theta))$
- Média: $\mu = \pi = \frac{\exp(\theta)}{1+\exp(\theta)}$ e $\nu(\mu) = \pi(1-\pi) = \frac{\exp(\theta)}{1+\exp(\theta)^2}$
- Observe que no modelo bernoulli, média e variância são relacionadas

$$\phi\nu(\mu) = \mu(1-\mu)$$

- Função de ligação natural: $\theta = \log(\mu/(1-\mu)$.

Função de Ligação Natural ou Canônica

$$g(\mu_i) = \eta_i = X_i'\beta = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

- Gaussiano: $g(\mu_i) = \eta_i$ (identidade)
- Bernoulli: $g(\mu_i) = \mathit{logit}(\eta_i)$.
- Poisson: $g(\mu_i) = \log(\eta_i)$

Inferência por MV

- Função de log-verossimilhança $logL(.) = l(.)$

$$L(\beta) = \prod_{i=1}^{N} f(y_i|\theta_i, \phi) = \prod_{i=1}^{N} \exp\{\phi^{-1}(y_i\theta_i - \psi(\theta_i)) + c(y_i, \phi)\}$$

- Equações escore: derivada de $l(.)$.
- Inferência baseada na teoria assintótica de MV.

Referências: Dobson (1990) e Cordeiro e Demétrio (201?)

Exemplo - Regressão Binária

- Uma amostra de 100 indivíduos acompanhados por um período de cinco anos.
- Resposta: ocorrência de doença coronariana.
- Resposta para cada indivíduo foi sim (1) ou não (0).
- Covariável de interesse: 8 faixas etárias (idade): 20-29, ..., 60-69.
- Aconteceram 43 ocorrências de doença coronariana.

Ref: Giolo (2010) pg. 98- Introdução à Análise de Dados Categóricos.

Entrada dos Dados

Existem duas formas de entrada dos dados para resposta binária.

- Forma 1: Uma linha para cada indivíduo:

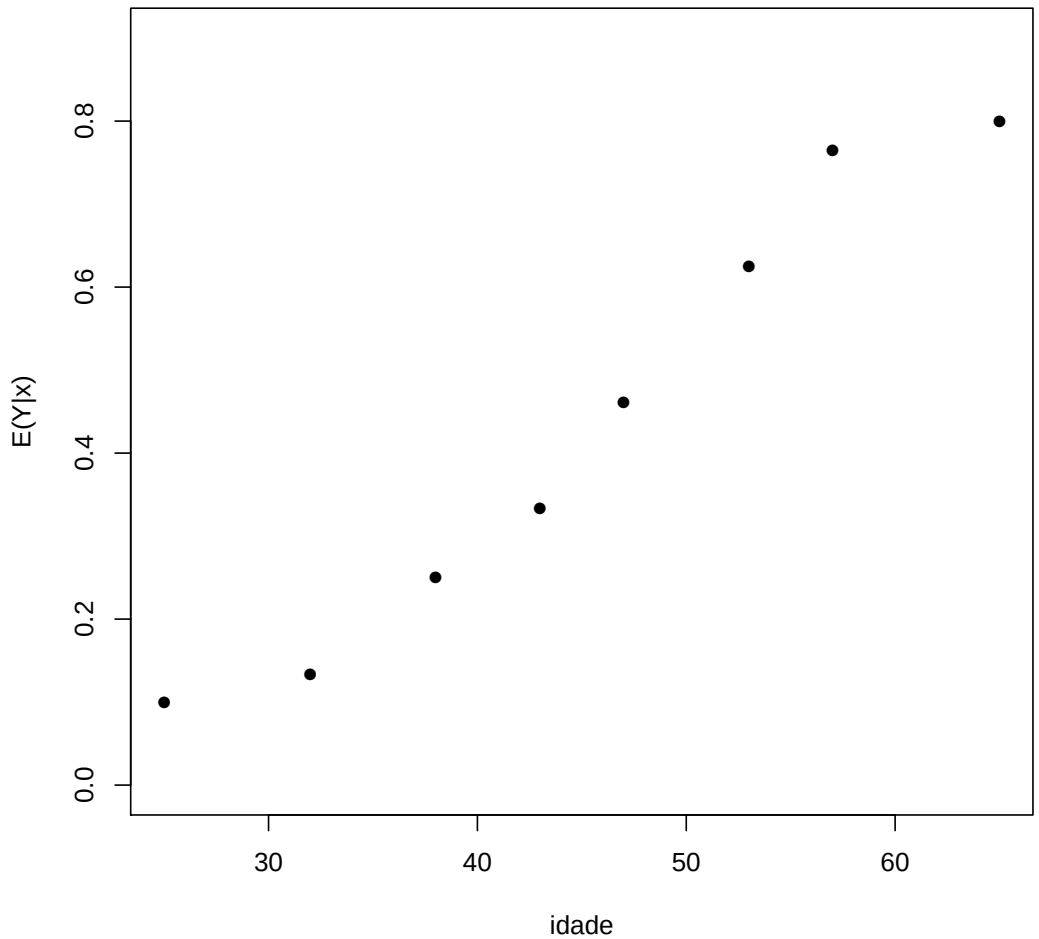
indivíduo	faixa etária	resposta
1	1 (25)	0
.....	..	.
100	5 (47)	1
Total	...	43

Entrada dos Dados

- Forma 2: Uma linha para cada combinação de covariáveis.

Faixa Etária	Sim	Não
20-29 (25)	1	9
30-34 (32)	2	13
35-39 (38)	3	9
40-44 (43)	5	10
45-49 (47)	6	7
50-54 (53)	5	3
55-59 (57)	13	4
60-69 (65)	8	2

Descrição Gráfica por Faixa Etária



$$logit(idade_i) = \log\{\mu_i/(1-\mu_i)\} = \beta_0 + \beta_1 idade_i$$

e

$$E(Y_i/idade_i) = P(Y_i = 1/idade_i)$$

O modelo logístico pode ser escrito como:

$$P(Y_i = 1/idade_i) = \frac{\exp(\beta_0 + \beta_1 idade_i)}{1 + \exp(\beta_0 + \beta_1 idade_i)}$$

Resultados do Ajuste MV

```
Coefficients:
             Estimate    Std. Error z value Pr(>|z|)
  (Intercept) -5.12300     1.11111  -4.611 4.01e-06 ***
   idade       0.10578     0.02337   4.527 5.99e-06 ***

Number of Fisher Scoring iterations: 4

> anova(ajust1,test="Chisq")

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
 NULL                     7    28.7015
 idade  1   28.118          6     0.5838 1.142e-07 ***
```

Resultados do Ajuste

Y: presença ou não de doença coronariana;
X: idade (em anos);
$n = 100$.

Variável	Estimativa	E.P.	Wald
Idade	0,106	0,023	4,53 ($p < 0,001$)
Constante	-5,123	1,11	-4,61 ($p < 0,001$)

$$\widehat{\pi}(x) = \frac{\exp(-5,12 + 0,106 \text{ idade})}{1 + \exp(-5,12 + 0,106 \text{ idade})}$$

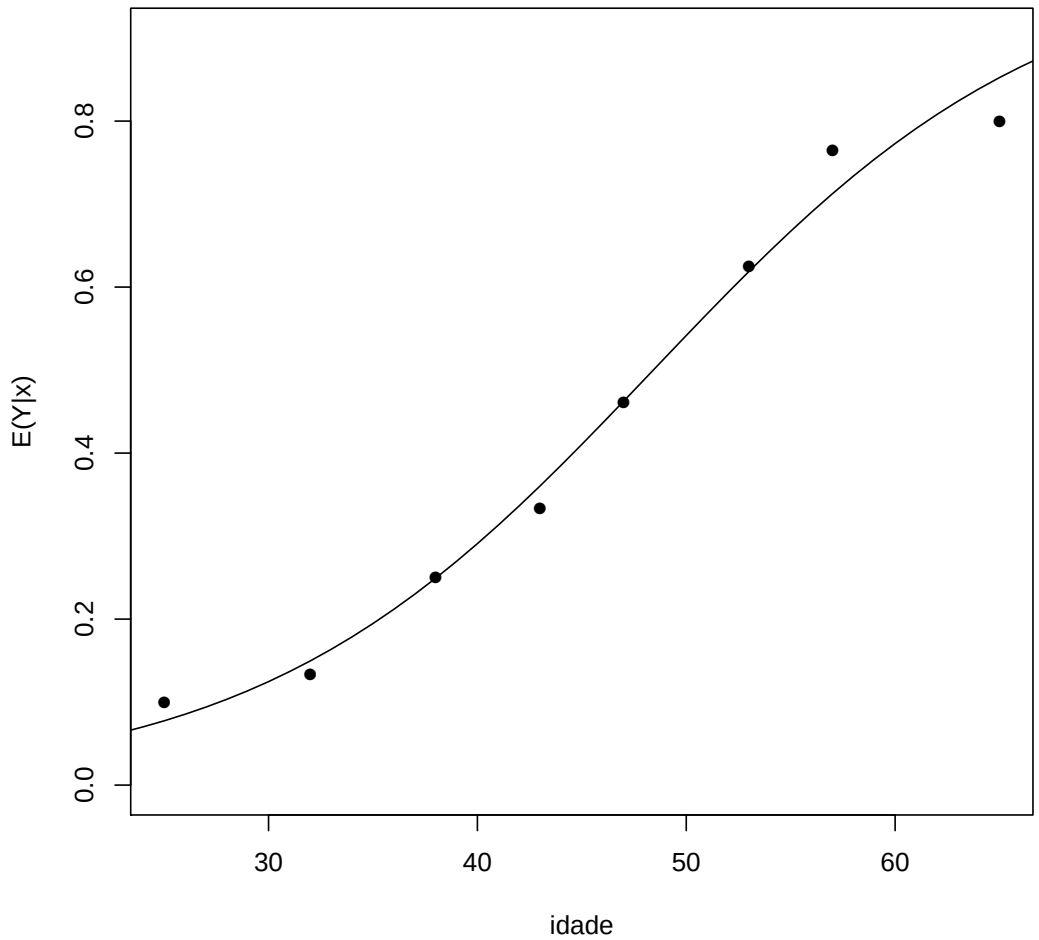
$$\widehat{logit}(x) = -5,12 + 0,106 \text{ idade}$$

$$\log(\text{verossimilhança}) = \log L(\widehat{\beta}_0, \widehat{\beta}_1) = -10,86$$

Sob $H_0 : \beta_1 = 0$, $logL(\widehat{\beta}_0) = -24,92$.

TRV = $2(-10,86 + 24,92) = \text{Null Deviance} - \text{Residual Deviance} = 28,118$.

Modelo Estimado



Interpretação dos Coeficientes

Interpretação: Razão de chances = $\exp(0,1058) = 1,11$ (1,06;1,16), isto significa que para o aumento de um ano na idade a chance de doença coronariana aumenta em 11%.

Outros MLG

- Y tem uma Bernoulli.
- Outras funções de ligação:
  - $\pi(x) = \Phi(x)$ (probit)
  - $\pi(x) = \exp - \exp(x)$ (complemento log-log)
  - etc (qualquer função de distribuição)

Modelos para Resposta Gaussiana Longitudinal

1. Modelo Marginal

$$Y_{ij} = X'_{ij}\beta + \varepsilon_{ij}$$

e

$$E(Y_{ij}|X_{ij}) = X'_{ij}\beta.$$

2. Modelo Condicional

$$Y_{ij} = X'_{ij}\beta + Z'_{ij}b_i + \varepsilon_{ij}$$

em que:

$(\beta)_{p\times 1}$: efeitos fixos;

$(b_i)_{q\times 1}$: efeitos aletaórios.

e,

$$b_i \sim N_q(0, \Sigma) \text{ e } \varepsilon_{ij} \sim N(0, \sigma^2)$$

Sendo b_i e ε_{ij} independentes.

- Média Condicional ou Específica por Indivíduo

$$E(Y_{ij}|b_i, X_{ij}) = X'_{ij}\beta + Z'_{ij}b_i.$$

e a Covariância Marginal

$$Var(Y_i) = Z_i \Sigma Z'_i + \sigma^2 I_{n_i}.$$

Modelos para Resposta Não-Gaussiana

1. $\mu_{ij} = E(Y_{ij}|X_{ij})$ (modelo marginal)

   $\mu_{ij} = E(Y_{ij}|b_i, X_{ij})$ (modelo condicional).

2. Modelo Bernoulli

   - Y_{ij} : 0/1 (Bernoulli)
   - função de ligação: logit (mais comum)

$$logit(\mu_{ij}) = X'_{ij}\beta \quad \text{Modelo Marginal}$$

$$logit(\mu_{ij}) = X'_{ij}\beta + Z'_{ij}b_i \quad \text{Modelo Condicional}$$

3 Modelo Poisson

- Y_{ij} :contagem (Poisson)
- função de ligação: logarítmica (mais comum)

$$\log(\mu_{ij}) = X'_{ij}\beta \quad \text{Modelo Marginal}$$

$$\log(\mu_{ij}) = X'_{ij}\beta + Z'_{ij}b_i \quad \text{Modelo Condicional}$$

Modelos Lineares Generalizados Longitudinais

1. Fácil transferência entre modelos (marginal e condicional) para resposta gaussiana.

2. Transferência díficil entre modelos quando a resposta não é gaussiana.

3. Modelos Marginais
   - Especificação completa: o ajuste por MV pode ser complicado.
   - Alternativa Não-Verossimilhança: MQG, GEE, etc.

4. Modelos Condicionais: ajuste complicado.

Resposta LOngitudinal Não-Gaussiano

1. Equações de Estimação Generalizadas

2. Modelos Lineares Mistos Generalizados

Modelos Marginais: GEE

Equações de Estimação Generalizadas

$$\sum_{i=1}^{N} D_i' V_i (Y_i - \mu_i) = 0,$$

em que

- $D_i = \partial \mu_i / \partial \beta$ e $\mu_i = g^{-1}(X_i \beta)$, ou seja, o inverso da função de ligação g.
- $$Var(Y_i) = V_i = \phi A_i^{1/2}(\beta) R_i(\alpha) A_i^{1/2}(\beta)$$

  em que A_i é uma matriz diagonal formada por $Var(Y_{ij})$, R_i é matriz de correlação de trabalho e ϕ é um parâmetro de dispersão/escala.
- $Var(\widehat{\beta})$ é estimada pela variância robusta (estimador sanduiche).

Formas de Correlação de Trabalho

- *independência*, $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{I}_k$;
  $\Rightarrow$ dados longitudinais não correlacionados.

- *simetria composta*, especifica que $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \rho 1_k 1_k' + (1-\rho)1_k$;
  $\Rightarrow$ mesma correlação.

- *AR-1*, para a qual $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \rho^{|j-j'|}$;
  $\Rightarrow$ válida para medidas igualmente espaçadas no tempo;

- *não estruturada* estima todas as $k(k-1)/2$ correlações de $\boldsymbol{R}$.

Variância do Estimador

1. *Naive* ou "baseada no modelo"

$$\widehat{Var}(\hat{\beta}) = \left(\sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \boldsymbol{R}_i(\hat{\alpha})^{-1} \hat{\boldsymbol{D}}_i \right)^{-1}. \tag{1}$$

2. *Robusta* ou "empírica"

$$\widehat{Var}(\hat{\beta}) = \boldsymbol{M}_0^{-1} \boldsymbol{M}_1 \boldsymbol{M}_0^{-1}, \tag{2}$$

em que

$$\begin{aligned} \boldsymbol{M}_0 &= \sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \boldsymbol{R}_i(\hat{\alpha})^{-1} \hat{\boldsymbol{D}}_i, \\ \boldsymbol{M}_1 &= \sum_{i=1}^{N} \hat{\boldsymbol{D}}_i' \boldsymbol{R}_i(\hat{\alpha})^{-1} (\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \hat{\boldsymbol{\mu}}_i)' \boldsymbol{R}_i(\hat{\alpha})^{-1} \hat{\boldsymbol{D}}_i. \end{aligned}$$

Exemplo: Bernoulli-logit

1. $\mu_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$.

2. $$logit(\mu_{ij}) = log(\mu_{ij}/(1 - \mu_{ij})) = X'_{ij}\beta$$

$$\mu_{ij} = \frac{e^{X'_{ij}\beta}}{1 + e^{X'_{ij}\beta}}$$

3. $$Var(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$$

4. $$\nu_{ij} = \mu_{ij}(1 - \mu_{ij}) \qquad A_i = diag(\nu_{i1}, \nu_{i2} \ldots, \nu_{in})$$

Exemplo: Poisson-log

1. $log(\mu_{ij}) = X'_{ij}\beta$.

$$\mu_{ij} = e^{X'_{ij}\beta}$$

2. $$Var(Y_{ij}) = \mu_{ij} = e^{X'_{ij}\beta}$$

3. $$\nu_{ij} = e^{X'_{ij}\beta} = \mu_{ij}$$

Estimando a Correlação de Trabalho

- Liang e Zeger (1986) utilizaram estimativas de momento para os parâmetros da matriz de correlação de trabalho.
- Ou seja, utilizar estimadores baseados nos resíduos para as quantidades envolvidas em R_i.
- Resíduos de Pearson:

$$e_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\nu}_{ij}}},$$

em que $\nu_{ij} = \mu_{ij}(1 - \mu_{ij})$ para resposta binária e $\nu_{ij} = \mu_{ij}$, para contagem.

Estimadores de Momento usando Resíduos

Estrutura	$Cor(Y_{ij}, Y_{il})$	Estimativa
Independência	0	-
Simetria Composta	α	$\hat{\alpha} = 1/N \sum_{i=1}^{N} 1/(n(n-1) \sum_{j \neq l} e_{ij} e_{il}$
AR1	α	$\hat{\alpha} = 1/N \sum_{i=1}^{N} 1/(n-1) \sum_{j \leq k-1} e_{ij} e_{ij+1}$
Não Estruturada	α_{jl}	$\hat{\alpha}_{jl} = 1/N \sum_{i=1}^{N} e_{ij} e_{il}$

$$\hat{\phi} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n} \sum_{i=1}^{n} e_{ij}^2$$

Ajustando GEE

1. Use MLE para encontrar a estimativa inicial para β (assumindo independência)
2. Encontre os resíduos e estime α e ϕ.
3. Faça iterações em 1-2 até a convergência.
4. Estime $Var(\hat{\beta})$ usando o estimador sanduíche.

Modelos Lineares Generalizados Mistos

1. Modelos Lineares Generalizados

   - Resposta na família exponencial: normal, gama, exponencial, Bernoulli, Poisson, etc.
   - Preditor Linear: $X_i'\beta$.
   - Função de Ligação: $g(\mu_i) = X_i'\beta$.

2. Modelos Lineares Generalizados Mistos

   Preditor Linear:

$$X_i\beta + Z_i b_i$$

.

Modelos Generalizado Misto Longitudinal

1. $$g(E(Y_{ij}|b_i)) = X'_{ij}\beta + Z'_{ij}b_i$$

em que:

$(\beta)_{p\times 1}$: efeitos fixos;

$(b_i)_{q\times 1}$: efeitos aletaórios.

e,

2. $$b_i \sim N_q(0, \Sigma) \text{ e } \varepsilon_{ij} \sim N(0, \sigma^2)$$

Sendo b_i e ε_{ij} independentes.

Função de Verossimilhança

$$
\begin{aligned}
L(\theta/y) &= \prod_{i=1}^{N} p(y_i/\theta) \\
&= \prod_{i=1}^{N} \int p(y_i, b_i/\theta) db_i \\
&= \prod_{i=1}^{N} \int p(y_i/b_i, \theta) p(b_i/\theta) db_i
\end{aligned}
$$

em que,

$$p(y_i/b_i, \theta) \sim \text{Bernoulli-logit/Poisson-log, etc}$$

e

$$p(b_i/\theta) \sim N_q(0, \Sigma)$$

Solução

- No modelo linear-normal, a integral pode ser resolvida analiticamente.
- Em geral, aproximações são necessárias no caso não-normal.
- Aproximação do integrando: Laplace
- Aproximação dos dados
- Aproximação da integral: quadratura gaussiana.

Usualmente, a combinação normal-logit não tem solução simples.

Interpretação dos Parâmetros

- Vetor β no GEE tem interpretação populacional. Ou seja, a mesma interpretação dos modelos transversais.
- Vetor β no modelo GLMM tem interpretação condicional sob o nível dos efeitos aleatórios. Ou seja, interpretação específica para cada indivíduo.
- Portanto, as estimativas dos modelos são diferentes.
- Em casos mais simples, os parâmetros apresentam uma relação.
- $$\frac{\hat{\beta}^{EA}}{\hat{\beta}^{M}} = \sqrt{c^2\sigma^2 + 1} > 1$$

Interpretação dos Parâmetros

Exemplo: Razão de Chances - Modelo Logit-normal

$$
\begin{aligned}
\text{RC} &= \frac{P(Y_i = 1|X_i = x+1)/P(Y_i = 0|X_i = x+1)}{P(Y_l = 1|X_l = x)/P(Y_l = 0|X_l = x)} \\
&= b_i + \beta(x+1) - (b_l + \beta x) \\
&= \beta + (b_i - b_l)
\end{aligned}
$$

Mecanismo Evacuatório de Récem-Nascidos

- 151 recém-nascidos acompanhados nos primeiros 12 meses de vida no Hospital das Clínicas da UFMG em 2010 e 2011.
- Acompanhamento mensal totalizando 1751 medidas (61 perdas)
- Resposta: Binárias: Dificuldade para evacuar.
- Variável temporal: idade (em dias ou meses).
- Covariáveis: 1- fixa (sexo) e 2- dependentes do tempo: Aleitamento materno, dieta (0/1): cereais; frutas; vegetais, carnes, etc.
- Objetivo: avaliar o comportamento temporal das respostas e seus respectivos indicadores.