Análise de Dados Longitudinais
Modelos Lineares Generalizados Longitudinais

Enrico A. Colosimo-UFMG
www.est.ufmg.br/∼enricoc

Respostas Longitudinal Não-Gaussiana

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

Exemplos

1. Mecanismo Evacuatório de Récem-Nascidos

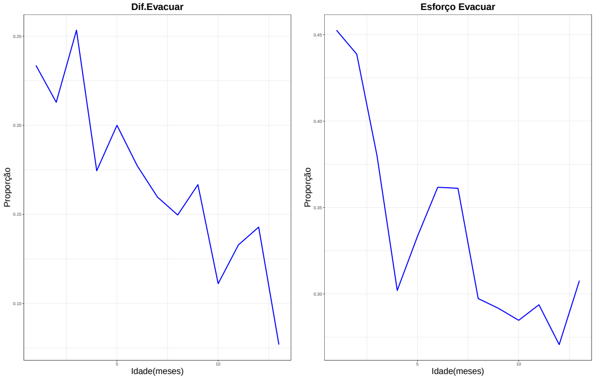
2. Fatores de Risco Coronariano: MCRF (Muscatine Coronary Risk Factor), (FLW, pag. 364)

Mecanismo Evacuatório de Récem-Nascidos

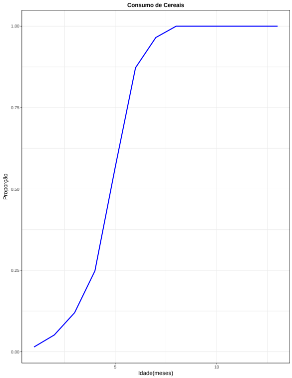
- 151 recém-nascidos foram acompanhados nos primeiros 12 meses de vida no Hospital das Clínicas da UFMG entre os anos de 2008 e 2009.
- Acompanhamento mensal totalizando 1751 medidas (61 perdas)
- Respostas: (1) Binárias: Dificuldade para evacuar, Esforço evacuatório, Dor ao evacuar e (2) Contagem: Freqüência 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 preditores.

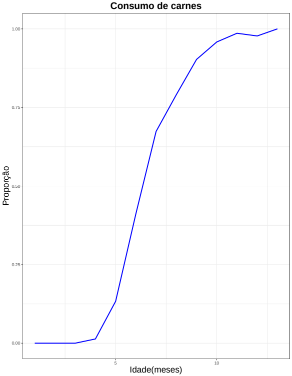
Resposta: Dificuldade e Esforço para Evacuar

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



Covariáveis: Consumo de Cereais e de Carnes





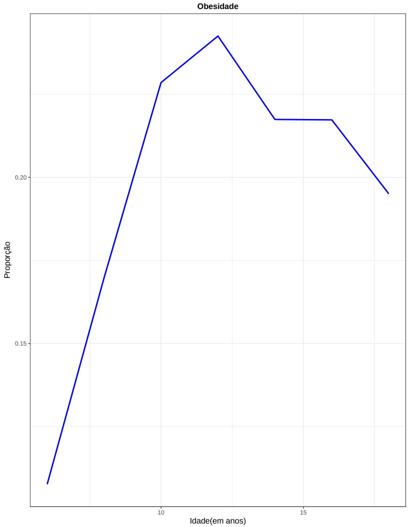
"Muscatine Coronary Risk Factor Study"

- 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).
- Resposta 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 de obesidade são os mesmos para meninos e meninas.

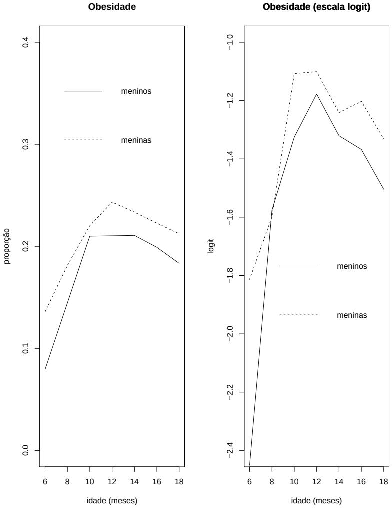
"Muscatine Coronary Risk Factor Study"

		Obesidade (%)		
Gênero	Coorte Idade	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

Resposta: Obesidade



Estratificação por Sexo



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.

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

- 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: $g(\mu_i) = logit(\eta_i)$.
- Poisson: $g(\mu_i) = \log(\eta_i)$

- 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 - Somente uma Medida

- 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:

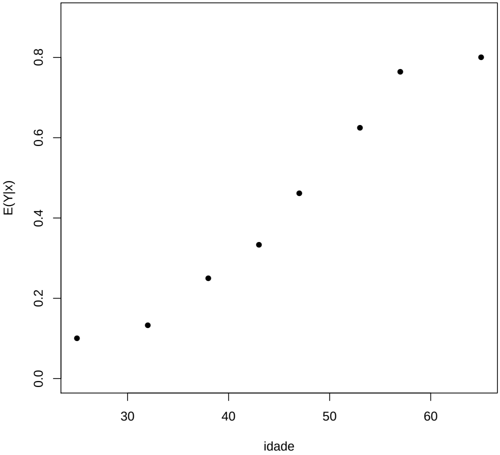
indivíduo	faixa etária	resposta
1	1 (25)	0
.....	..	.
100	5 (47)	1
Total	...	43

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) -5.25903    1.13739  -4.624 3.77e-06 ***
idade        0.10904    0.02394   4.554 5.25e-06 ***

Number of Fisher Scoring iterations: 4

> anova(ajust1,test="Chisq")

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       7    28.7015
idade  1   28.303          6     0.3984 1.037e-07 ***
```

Resultados do Ajuste

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

Variável	Estimativa	E.P.	Wald
Idade	0,109	0,023	4,55 ($p < 0,001$)
Constante	-5,259	1,13	-4,62 ($p < 0,001$)

$$\widehat{\pi}(x) = \frac{\exp(-5,25 + 0,109 \text{ idade})}{1 + \exp(-5,25 + 0,109 \text{ idade})}$$

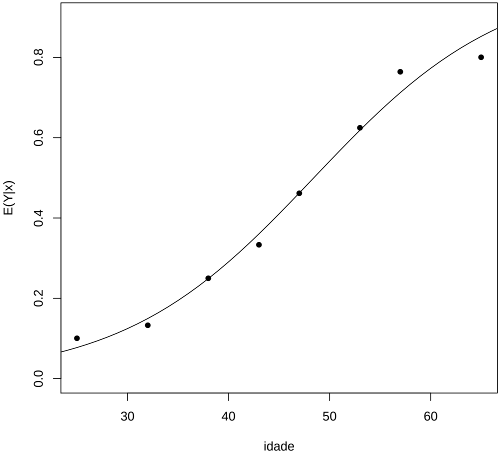
$$\widehat{logit}(x) = -5,25 + 0,109 \text{ idade}$$

$$\log(\text{verossimilhança}) = \log L(\widehat{\beta}_0, \widehat{\beta}_1) = -10,76$$

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

TRV = $2(-10,76 + 24,91) =$ Null Deviance $-$ Residual Deviance $= 28,303$.

Modelo Estimado



Interpretação dos Coeficientes

Interpretação: Razão de chances = $\exp(0,109)$ = 1,11 (1,06;1,17), isto significa que para o aumento de um ano na idade a chance de doença coronariana aumenta em 11%.

Outras Funções de Ligação

- 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)

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 aleatórios.

e,

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

Sendo b_i e ε_{ij} independentes.

Modelos para Resposta Gaussiana

- Média Condicional ou Específica por Indivíduo

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

e a Covariância Marginal

$$Var(Y_i) = Z_i \Sigma Z_i' + \sigma^2 I_n.$$

Modelos para Resposta Não-Gaussiana

- Extensão natural dos modelos lineares generalizados;
- Levar em consideração a variação intra-indivíduo;
- Duas classes de modelos:
  - Modelos Marginais: baseado na média da população;
  - Modelos de Efeitos Mistos: específico por indivíduo
- Interpretação dos coeficientes de regressão para os dois modelos:
  - Mesma para resposta contínua (gaussiana);
  - Diferente para resposta bínaria, poisson
- Escolha do modelo deve ser baseada nos objetivos do estudo.

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)

$$\log(\mu_{ij}) = X_{ij}'\beta \qquad \text{Modelo Marginal}$$

$$\log(\mu_{ij}) = X_{ij}'\beta + Z_{ij}'b_i \qquad \text{Modelo Condicional}$$

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.

1. Equações de Estimação Generalizadas (Modelo Marginal)

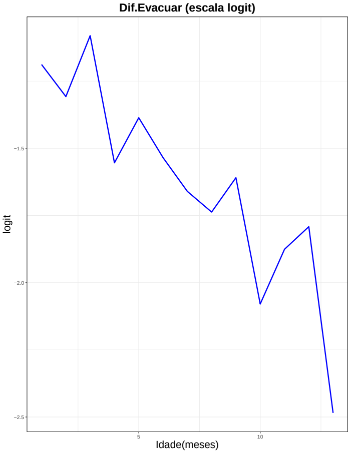
2. Modelos Lineares Mistos Generalizados (Modelo de Efeitos Aleatórios)

Modelando a Média

- Necessitamos modelar a média em ambos modelos;
- Usamos Análise Exploratória para propor uma forma paramétrica;
- Precisamos fazer os gráficos na escala da função de ligação.

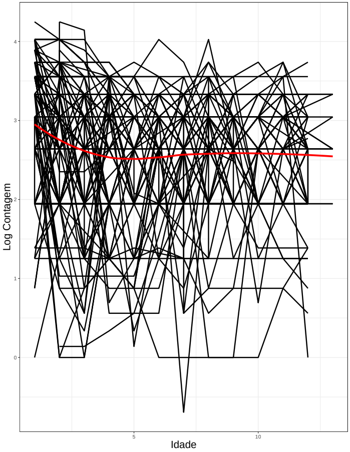
Resposta Binária: Dificuldade para Evacuar

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



Resposta Contagem: Freq. Evac / semana

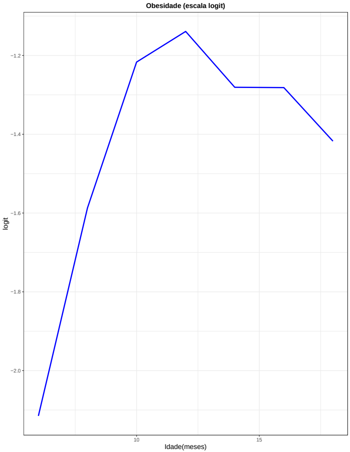
Obs.: idade foi arrendonda para mês (um único digito).



"Muscatine Coronary Risk Factor Study"

		Obesidade (%)		
Gênero	Coorte Idade	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

Resposta binária: obesidade - Escala logit



Modelos Marginais: GEE (Liang e Zeger, 1986)

- Modelo Marginal significa que a média (populacional) depende somente das covariáveis de interesse.

- Em outras palavras, a média não incorpora dependência através de efeitos aleatórios nem de repostas em tempos anteriores.

- Não é necessário assumir distribuição conjunta para a resposta Y_i: GEE.

- Não é necessário ser balanceado.

- Assume que a distribuição marginal pertence a classe dos modelos lineares generalizados.

Equações de Estimação Generalizadas

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

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

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

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

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

Variância do Estimador

1. *Naive* ou "baseada no modelo"

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

2. *Robusta* ou "empírica"

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

em que

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

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.

Teste de hipóteses - GEE

1. Não existe TRV pois não temos função de verossimllhaça.
2. Testes e Intervalos de confiança são baseados na estatística de Wald
3. $$H_0 : C\beta = 0 \;\; vs H_1 : C\beta =\neq 0$$

   para C de posto $q < p$.

   $$Q = (C\widehat{\beta})'(C\widehat{Var}(\hat{\beta})C')^{-1}(C\widehat{\beta})$$

   Q sob H_0 tem uma distribuição qui-quadrado limite com q graus de liberdade.
4. o comando anova no pacote geepack faz este teste.

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.

Resultados do Ajuste GEE - Resposta binária

```
geeglm(formula = difpevac ~ sexo + idademeses + aleitamcod3,
    family = binomial(link = "logit"), data = dados, id = Paciente,
    corstr = "exchangeable", std.err = "san.se")

 Coefficients:
             Estimate Std.err  Wald Pr(>|W|)
(Intercept)   -0.9275  0.2182 18.07  2.1e-05 ***
sexo1         -0.4858  0.2287  4.51  0.03362 *
idademeses    -0.1240  0.0338 13.49  0.00024 ***
aleitamcod31   0.2286  0.2300  0.99  0.32030
aleitamcod32   0.7969  0.3037  6.88  0.00870 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     1.02   0.189

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.207   0.058
Number of clusters:   151   Maximum cluster size: 12
```

Resultados do Ajuste

Y: (0=não /1=sim) Dificuldade para evacuar

X: Covariáveis: (1) idade em meses: (2) sexo (0-menina e 1- menino), (3) consumo de cereais (0-não e 1-sim), (4) consumo de carne (0-não e 1-sim) e (5) aleitamento materno (0-peito, 1-misto e 2-artificial)).

n = 151 pacientes e 1751 medições.

Variável	Estimativa	E.P.	Wald
Idade	-0,124	0,038	13,49 ($p < 0,001$)
Sexo(Menino)	-0,485	0,22	4,51 ($p = 0,033$)
Aleitamento(Misto)	0,228	0,23	0,99 ($p = 0,32$)
Aleitamento(Artificial)	0,796	0,30	6,88 ($p = 0,008$)
Constante	-0,927	0,218	

Interpretação dos Coeficientes

- Idade: Razão de chances = $\exp(-0,124)$ = 0,83 (0,81;0,95), isto significa que com o aumento de um ano na idade a chance de dificuldade em evacuar reduz em 17%;
- Sexo: Razão de chances = $\exp(-0,485)$ = 0,615 (0,40;0,947), isto significa que a chance de meninos terem dificuldade em evacuar é 39% menor do a chance das meninas.
- Aleitamento: Razão de chances = $\exp(0,796)$ = 2,22 (1,23;3,99), isto significa que a chance de récem nascidos que tiveram aleitamento artificial ter dificuldade em evacuar é 2,2 vezes a chance daqueles que tiveram aleitamento no peito.

Ajustando GEE-Poisson

```
Call:
geeglm(formula = freqevac ~ idademeses + aleitamcod3 + carnes,
    family = poisson(link = "log"), data = dados, id = Paciente,
    corstr = "exchangeable", std.err = "san.se")

 Coefficients:
             Estimate  Std.err    Wald Pr(>|W|)
(Intercept)   3.16227  0.05633 3151.97  < 2e-16 ***
idademeses   -0.03185  0.00737   18.67  1.6e-05 ***
aleitamcod31 -0.31691  0.05209   37.02  1.2e-09 ***
aleitamcod32 -0.19479  0.05925   10.81    0.001 **
carnes1       0.12056  0.04812    6.28    0.012 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Estimated Scale Parameters:
            Estimate Std.err
(Intercept)     5.59   0.399

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.303   0.039
Number of clusters:   151   Maximum cluster size: 12
```

Resultados do Ajuste - Resposta Contagem - Modelo Poisson

Y: Freqüência evacuatória (número de evacuações/semana)

X: Covariáveis: (1) idade em meses: (2) sexo (0-menina e 1- menino), (3) consumo de cereais (0-não e 1-sim), (4) consumo de carne (0-não e 1-sim) e (5) aleitamento materno(0-peito, 1-misto e 2-artificial)).

n = 151 pacientes e 1751 medições.

Variável	Estimativa	E.P.	Wald
Idade	$-$0,031	0,007	18,67 ($p < 0,001$)
Carne(Sim)	0,12	0,04	6,28 ($p = 0,012$)
Aleitamento(Misto)	-0,316	0,052	37,02 ($p < 0,001$)
Aleitamento(Artificial)	-0,194	0,048	10,81 ($p = 0,001$)
Constante	3,16	0,056	

Interpretação dos Coeficientes

- Idade: Razão de frequências evacuatórias = $\exp(-0,031)$ = 0,96 (0,95;0,98), isto significa que com o aumento de um ano na idade a frequência evacuatória semanal diminui em 4%.
- Carne: Razão de frequências evacuatórias = $\exp(0,12)$ = 1,12 (1,04;1,22), isto significa que a frequência evacuatória semanal em recém nascidos que consomem carne é 12% maior do que aqueles que não consomem carne em sua dieta.
- Aleitamento materno: Razão de frequências evacuatórias = $\exp(-0,316)$ = 0,73 (0,658;0,807), isto significa que a frequência evacuatória semanal de récem nascidos que tiveram aleitamento misto é 27% menor do que aqueles que foram amamentados no peito.
- Razão de frequências evacuatórias = $\exp(-0,194)$ = 0,82 (0,749;0,904), isto significa que a frequência evacuatória semanal de récem nascidos que tiveram aleitamento artificial é 18% menor do que aqueles que tiveram amamentamento no peito.

Modelos Lineares Generalizados Mistos

1. GEE
   - Média e Covariância são modeladas separadamente.
   - Interpretação dos coeficientes é a nível populacional.
2. Modelos Lineares Generalizados Mistos
   - Forma alternativa de tratar a associação intra-indivíduo é introduzindo efeitos aleatórios.
   - O efeito aleatório reflete uma heterogeneidade entre indivíduos e induz uma estrutura de correlação para as medidas repetidas.
   - Produz dificuldades conceituais (interpretação) e analíticas. Interpretações especificas de indivíduos/clusters.

Modelos Lineares Generalizados Mistos

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

   Preditor Linear:

$$\eta_{ij} = X_{ij}'\beta + Z_{ij}'b_i,$$

   em que:

   $(\beta)_{p\times 1}$: efeitos fixos;
   $(b_i)_{q\times 1}$: efeitos aletaórios.

Modelos Linear Misto Longitudinal

- Modelo Linear Misto:

$$Y_{ij} = X_{ij}'\beta + Z_{ij}'b_i + \varepsilon_{ij}$$

ou

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i$$

- Média Condicional (Preditor Linear)

$$E(Y_{ij}|b_i) = X_{ij}'\beta + Z_i'b_i$$

.

- Média Populacional (marginal)

$$E(Y_{ij}) = X_i\beta$$

.

Modelos Generalizado Misto Longitudinal

1. Família Exponencial

   - Distribuição condicional Y_{ij}/b_i pertence a família exponencial.
   - Dado os efeitos aleatórios, Y_{ij} e $Y_{ij'}$, para $j \neq j'$, são independentes (independência condicional).

2. Função de Ligação

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

3. Distribuição dos Efeitos Aleatórios

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

Sendo b_i e ε_{ij} independentes.

Modelos Generalizado Misto para Respostas Binárias

- Distribuição condicional Y_{ij}/b_i tem distribuição Bernoulli com:
  - $E(Y_{ij}/b_i) = P(Y_{ij} = 1/b_i)$;
  - $Var(Y_{ij}/b_i) = E(Y_{ij}/b_i)(1 - E(Y_{ij}/b_i))$;
  - $\phi = 1$.
- Função de Ligação (logit)

$$g(E(Y_{ij}|b_i)) = log(P(Y_{ij} = 1/b_i)/P(Y_{ij} = 0/b_i) = X'_{ij}\beta + Z'_{ij}b_i$$

- Distribuição condicional Y_{ij}/b_i tem distribuição de Poissson com:
  - $Var(Y_{ij}/b_i) = E(Y_{ij}/b_i)$;
  - $\phi = 1$.
- Função de Ligação (logit)

$$g(E(Y_{ij}|b_i)) = log(E(Y_{ij}/b_i)) = X_{ij}'\beta + Z_{ij}'b_i$$

Função de Verossimilhança

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

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 do Modelo

- Os parâmetros β não tem interpretação populacional, como aqueles do modelo marginal.
- Os parâmetros β têm interpretação específica de indivíduo/sujeito.
- Ou seja, eles representam o efeito de covariáveis na resposta média de um especifico indivíduo.
- Por exemplo, na regressão logística com efeito aleatório no intercepto e somente um regressor x:

$$\log(P(Y_{ij} = 1/b_i, X = x)/P(Y_{ij} = 0/b_i, X = x)) = b_i + \beta x$$

Interpretação dos Parâmetros do Modelo

- No modelo linear misto, as interpretações populacional (marginal) e especifica (condicional) são equivalentes.

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

- Nos modelos lineares generalizados em que a função de ligação não é a identidade (não-linear) esta relação não vale.

  Ou seja,

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

$$g(E(Y_{ij}|b_i)) \neq X'_{ij}\beta$$

Interpretação dos Parâmetros do Modelo

- De forma a termos a interpretação usual de $RC = e^{\beta}$ para o aumento em uma unidade de X, devemos cancelar o efeito aleatório b_i.
- Daí, β é o log odds da resposta por aumento de uma unidade em X, para qualquer indivíduo tendo uma propensão de resposta positiva *bi*.
- Coeficientes de regressão especifico por indivíduos. MLGM são, portanto, mais úteis quando o objetivo científico é fazer inferência em indivíduos ao invés de médias populacionais.
- Tais interpretações ficam sem sentido se as covariáveis forem entre indivíduos, por exemplo, gênero ou tratamento.

Interpretação dos Parâmetros

- Portanto, as estimativas dos modelos são diferentes.
- Em casos mais simples, os parâmetros apresentam uma relação.
- Por exemplo, modelo logístico com somente intercepto aleatório,

$$\frac{\hat{\beta}^{EA}}{\hat{\beta}^{M}} = \sqrt{c^2\sigma^2_{\beta_0} + 1} > 1$$

em que

$$c = \frac{16\sqrt(3)}{15\pi} = 0,346$$

Interpretação dos Parâmetros

Exemplo: Razão de Chances - Modelo Logit-normal

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

Ajuste do Modelo

- Parte numérica complicada pois é necessário encontrar um máximo da função de verossimilhança que envolve uma integral.
- Em particular, a função de ligação logística e a distribuição normal não possuem boa relação.
- O pacote lme4 do R é o usualmente utilizado para ajuste dos MLGM.
- A função lmer é usada para este propóstito.

```
Basic syntax,
lmer(formula, data, family=NULL, REML=TRUE)
```

- Exemplos:
  - Random intercept lmer(Y age + (1| Id), data=dados)
  - Random intercept and slope lmer(Y age + (age | id), data=dados)

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.

Ajustando Modelo Misto - Resposta binária

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: difpevac ~ (1 | Paciente) + aleitamcod3 + sexo * idademeses
   Data: dados

     AIC      BIC   logLik deviance df.resid
    1427     1465     -706     1413     1743

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.050 -0.412 -0.254 -0.169  4.315

Random effects:
 Groups   Name        Variance Std.Dev.
 Paciente (Intercept) 1.89     1.38
Number of obs: 1750, groups:  Paciente, 151

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.9897     0.2665   -3.71  0.00020 ***
aleitamcod31       0.3089     0.2335    1.32  0.18587
aleitamcod32       1.0345     0.3138    3.30  0.00098 ***
sexo1             -1.2812     0.3867   -3.31  0.00092 ***
idademeses        -0.2085     0.0353   -5.92  3.3e-09 ***
sexo1:idademeses   0.1155     0.0437    2.64  0.00823 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Resultados do Ajuste

Y: (0=não /1=sim) Dificuldade para evacuar

X: Covariáveis: (1) idade em meses: (2) sexo (0-menina e 1- menino), (3) consumo de cereais (0-não e 1-sim), (4) consumo de carne (0-não e 1-sim) e (5) aleitamento materno (0-peito, 1-misto e 2-artificial)).

n = 151 pacientes e 1751 medições.

Parte fixa:

Variável	Estimativa	E.P.	Z
Idade	-0,208	0,03	-5,92 ($p < 0,001$)
Sexo(Menino)	-1,281	0,38	-3,31 ($p < 0,001$)
Aleitamento(Misto)	0,308	0,23	1,32 ($p = 0,18$)
Aleitamento(Artificial)	1,034	0,31	3,30 ($p < 0,001$)
Idade*Sexo(Masculino)	0,115	0,04	2,64 ($p = 0,008$)
Constante	-0,927	0,218	18,07 ($p < 0,001$)

Variância estimada da parte aleatória é 1,89 e o desvio padrão correspondente é 1,38.

Ajustando Modelo Misto-Poisson

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: poisson  ( log )
Formula: freqevac ~ (1 | Paciente) + sexo + idademeses + cereais + aleitamcod3 +      carnes
   Data: dados
     AIC      BIC   logLik deviance df.resid
     Inf      Inf     -Inf      Inf     1743
Random effects:
 Groups   Name        Std.Dev.
 Paciente (Intercept) 1
Number of obs: 1751, groups:  Paciente, 151
Fixed Effects:
 (Intercept)         sexo1    idademeses      cereais1  aleitamcod31  aleitamcod32       carnes1
      3.0092        0.2066       -0.0366        0.1083       -0.3674       -0.2381        0.0901
convergence code 0; 1 optimizer warnings; 13000 lme4 warnings
```

Resultados do Ajuste - Modelo Misto-Binomial Negativo

Y: Freqüência evacuatória (número de evacuações/semana)

X: Covariáveis: (1) idade em meses: (2) sexo (0-menina e 1- menino), (3) consumo de cereais (0-não e 1-sim), (4) consumo de carne (0-não e 1-sim) e (5) aleitamento materno(0-peito, 1-misto e 2-artificial)).

n = 151 pacientes e 1751 medições.

Parte fixa:

Variável	Estimativa	E.P.	Z
Idade	-0,029	0,006	-4,91 ($p < 0,001$)
Carnes(Sim)	0,136	0,043	3,11 ($p = 0,001$)
Aleitamento(Misto)	-0,316	0,039	-7,96 ($p < 0,001$)
Aleitamento(Artificial)	-0,198	0,056	-3,51 ($p < 0,001$)
Constante	3,058	0,049	62,18 ($p < 0,001$)

Variância estimada da parte aleatória é 0,106 e o desvio padrão correspondente é 0,325.

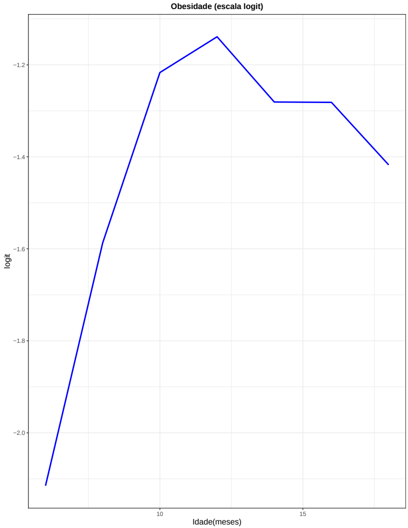
Ajustando Modelo Misto-Binomial Negativo

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: Negative Binomial(5.54)  ( log )
Formula: freqevac ~ (1 | Paciente) + sexo + idademeses + aleitamcod3 +      carnes
   Data: dados

     AIC      BIC   logLik deviance df.resid
   12237    12281    -6110    12221     1743

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.013 -0.672 -0.057  0.499  5.770

Random effects:
 Groups   Name        Variance Std.Dev.
 Paciente (Intercept) 0.106    0.325
Number of obs: 1751, groups:  Paciente, 151

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.05817    0.04918   62.18  < 2e-16 ***
sexo1         0.04932    0.05766    0.86  0.39238
idademeses   -0.02922    0.00694   -4.21  2.6e-05 ***
aleitamcod31 -0.31619    0.03973   -7.96  1.7e-15 ***
aleitamcod32 -0.19843    0.05651   -3.51  0.00045 ***
carnes1       0.13620    0.04385    3.11  0.00189 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estudo de Caso: "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).
- Resposta binária: obesidade (sim/não).
- 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 de obesidade são os mesmos para meninos e meninas.

Resposta: Obesidade na escala logit



Modelo Completo: efeito de interação e coorte

$$\begin{aligned}
\text{logit}(X_{ij}) &= \beta_1 + \beta_2 \text{sexo}_i + \beta_3 \text{idade}_{i1} + \beta_4 \text{idade}_{i1}^2 \\
&+ \beta_5 \text{sexo}_i \times \text{idade}_{i1} + \beta_6 \text{sexo}_i \times \text{idade}_{i1}^2 \\
&+ \beta_7 (\text{idade}_{ij} - \text{idade}_{i1}) + \beta_8 (\text{idade}_{ij}^2 - \text{idade}_{i1}^2) \\
&+ \beta_9 \text{sexo}_i \times (\text{idade}_{ij} - \text{idade}_{i1}) + \beta_{10} \\
&+ \text{sexo}_i \times (\text{idade}_{ij}^2 - \text{idade}_{i1}^2)
\end{aligned}$$

Teste para efeito de coorte

$$H_0 : \beta_3 - \beta_7 = \beta_4 - \beta_8 = \beta_5 - \beta_9 = \beta_6 - \beta_{10} = 0$$

```
>   anova(out0,out)
Analysis of 'Wald statistic' Table

Model 1 Y ~ GENDER + age1 + I(age1^2) + GENDER:age1 + GENDER:I(age1^2) + I(AGE - age1) + I(AGE^2 -
Model 2 Y ~ AGE + I(AGE^2) + GENDER + AGE:GENDER + I(AGE^2):GENDER
  Df   X2 P(>|Chi|)
1  4 2.02      0.73
```

Modelo - Sem efeito de coorte

$$
\begin{aligned}
\text{logit}(X_{ij}) &= \beta_1 + \beta_2 \text{sexo}_i + \beta_3 \text{idade}_{ij} + \beta_4 \text{idade}_{ij}^2 \\
&+ \beta_5 \text{sexo}_i \times \text{idade}_{ij} + \beta_6 \text{sexo}_i \times \text{idade}_{ij}^2
\end{aligned}
$$

Teste para efeito de interação

$$H_0 : \beta_5 = \beta_6 = 0$$

```
>  anova(out,out1)

Analysis of 'Wald statistic' Table

Model 1 Y ~ AGE + I(AGE^2) + GENDER + AGE:GENDER + I(AGE^2):GENDER
Model 2 Y ~ AGE + GENDER + I(AGE^2)
  Df   X2 P(>|Chi|)
1  2 0.96      0.62
```

Resultados do Ajuste GEE - Resposta binária obesidade

```
geeglm(formula = Y ~ AGE + I(AGE^2) + GENDER, family = binomial(link = "logit"),
    data = dados, id = ID, corstr = "exchangeable", std.err = "san.se")

 Coefficients:
            Estimate  Std.err  Wald Pr(>|W|)
(Intercept) -3.98900  0.33954 138.0  < 2e-16 ***
AGE          0.41856  0.05632  55.2  1.1e-13 ***
I(AGE^2)    -0.01570  0.00231  46.3  1.0e-11 ***
GENDER       0.14713  0.06271   5.5    0.019 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    0.991  0.0271
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.543  0.0209
Number of clusters:   4856  Maximum cluster size: 3
```

Resultados do Ajuste

Y: (0=não/1=sim) Obesidade

X: Covariáveis: idade em anos:(2) sexo (0-menino e 1- menina). n = 4856 pacientes e 14568 medições.

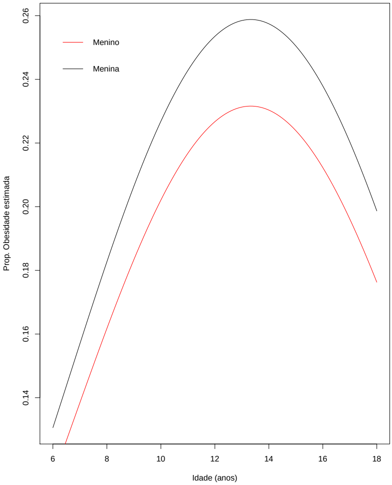
Variável	Estimativa	E.P.	Wald
Idade	0.419	0,056	55.2 ($p < 0,001$)
Idade^2	-0.016	0.0023	46.3 ($p < 0,001$)
Sexo(Menina)	0,147	0,063	5,5 ($p = 0,019$)

Interpretação dos Coeficientes

- Idade: Razão de chances muda de idade
    - mudando de 9 para 10 anos;
      RC = $\exp(2.62 - 2.5) = 1,13$, aumento em 13% a obesidade,
    - mudando de 13 para 14 anos
      RC = 1, não tem mudança.
    - mudando de 15 para 16 anos;
      RC = $\exp(2.62 - 2.75) = 0,88$, redução em 12% a obesidade,

- Sexo: Razão de chances = $\exp(-0,150)$ = 1,16 (1,03;1,31), isto significa que a chance de meninnas serem obesas é 16% maior que a chance dos meninos.

Resposta: Obesidade



Ajustando Modelo Misto - Resposta binária: obesidade

- Modelo misto tem interpretação especifica por indivíduo.
- Frequentemente, existem problemas numéricos no ajuste dos modelos mistos, especialmente o binário com ligação logit.
- Componente fixo em idade apresenta um comportamento não-linear. Vamos continuar utilizando o termo quadrático. Vamos centrar idade na média para evitar o efeito de colinearidade, que torna mais sério o problema numéricos.
- Iremos utilizar duas estruturas de efeitos aleatórios: (1) somente no intercepto; (2) intercepto e inclinação.

Ajustando Modelo Misto - Resposta binária: obesidade

Idade foi centrada na média

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: Y ~ (idadec | ID) + idadec + I(idadec^2) + GENDER
   Data: dados

     AIC      BIC   logLik deviance df.resid
    7347     7397    -3666     7333     9849

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.6212 -0.0087 -0.0076 -0.0014  1.2443

Random effects:
 Groups Name        Variance Std.Dev. Corr
 ID     (Intercept) 309.5    17.59
        idadec       25.4     5.04    0.18
Number of obs: 9856, groups:  ID, 4856

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.1841     0.2621  -35.04  < 2e-16 ***
idadec        0.0447     0.0768    0.58     0.56
I(idadec^2)  -0.1156     0.0188   -6.15  7.6e-10 ***
GENDER        0.1173     0.2540    0.46     0.64
---
```

Resultados do Ajuste

Y: (0=não/1=sim) Obesidade

X: Covariáveis: (1) idade em anos e (2) sexo (0-menino e 1- menina)

n = 4856 pacientes e 14568 medições.

Somente idade foi importante!