Especialização em *Data Science* e Estatística Aplicada

Módulo IV – Análise de Dados Longitudinais

Prof. Dr. Márcio Luis Lanfredi Viola

Goiânia, 2025

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN FACULDADE DE ENFERMAGEM
UFG UNIVERSIDADE FEDERAL DE GOIÁS

Conteúdo Programático

- Conceitos Básicos.
- Principais delineamentos para dados longitudinais.
- Modelos lineares gerais para dados longitudinais.
- Modelos lineares generalizados para dados longitudinais.
- Modelos com efeitos aleatórios.
- Aplicações.

Conteúdo - Aula 4

- Modelos lineares generalizados para dados longitudinais.
- Equações de estimação generalizadas.
- Exemplos.

Família Exponencial Linear

Uma distribuição pertence à família exponencial linear se a sua função de probabilidade (ou função densidade) puder ser expressa na forma

$$f(y|\theta) = \exp[\phi\,[y\theta - b(\theta)] + c(y, \phi)],$$

em que b e c são funções conhecidas e ϕ é um parâmetro de escala.

Exemplo 1

Considere Y ~ Bernoulli(p). Sua função de probabilidade é

$$P(Y = y) = p^y(1-p)^{1-y},\ y \in \{0,1\},$$

que pode ser expressa na forma da família exponencial linear:

$$P(Y = y) = p^y(1-p)^{1-y} = \exp[\ln(p^y(1-p)^{1-y})] = \exp[\ln(p^y) + \ln(1-p)^{1-y})] =$$

$$\exp[y \ln(p) + (1-y) \ln(1-p)] = \exp[y \ln(p/(1-p)) + \ln(1-p)] = \exp[y\theta - b(\theta)],$$

com $\phi = 1$, $\theta = \ln(p/(1-p))$, $b(\theta) = -\ln(1-p)$ e $c(y, \phi) = 1$.

Além da distribuição Bernoulli, que é usada para modelar dados binários, há outras distribuições que também pertencem à família exponencial linear como, por exemplo:

- Distribuição Binomial: usada para modelar dados binários agregados;
- Distribuição Poisson: usada para modelar dados de contagem;
- Distribuição Normal: utilizada para modelar dados contínuos e simétricos;
- Distribuição Gama: utilizada para modelar dados contínuos, positivos e assimétricos;
- Distribuição Gaussiana Inversa: usada para modelar dados contínuos, positivos e assimétricos.

Propriedades

Se a distribuição de Y pertencer à família exponencial linear, então

$$\mathrm{E}(Y) = \mu = \frac{db(\theta)}{d\theta},$$

e

$$\mathrm{Var}(Y) = \phi^{-1}\frac{d^2 b(\theta)}{d\theta^2}.$$

Modelo Linear Generalizado

Um Modelo Linear Generalizado (MLG) possui três componentes:

(1) **Componente aleatório:** Representado por um conjunto de variáveis aleatórias independentes Y_1, Y_2, ..., Y_n associadas a uma distribuição pertencente à família exponencial linear com parâmetros θ_i, i =1, ..., n, e ϕ ;

(2) **Componente sistemático:** Este componente engloba as covariáveis e é representado como $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, em que $\boldsymbol{x}_i$ é o vetor correspondente à i-ésima observação das covariáveis e $\boldsymbol{\beta}$ é o vetor associado aos parâmetros do modelo;

(3) **Função de ligação:** É uma função estritamente monótona e duplamente diferenciável, g, que relaciona o componente aleatório ao sistemático.

Modelo Linear Generalizado

Um modelo linear generalizado é formulado como

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

para $i = 1, \ldots, n$.

Modelo Linear Generalizado

Ajustar um MLG consiste em:

- Escolher uma distribuição adequada para a variável resposta;
- Escolher as variáveis explicativas (covariáveis) que entrarão no modelo;
- Escolher uma função de ligação.

Exemplo 2: Modelo de Regressão Logística

- A Regressão Logística é um dos modelos usados para a modelagem de dados binários, ou seja, uma variável resposta que possui distribuição Bernoulli;

- Este modelo relaciona a probabilidade de sucesso da variável resposta *Y* (P(*Y*=1)) com as variáveis explicativas (covariáveis) por meio da função de ligação logito, expressa por

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = \ln\left(\frac{p}{1-p}\right),$$

em que $p = P(Y = 1)$ e $\mu = E(Y) = p$.

Observação

A função de ligação logito é expressa em termos de *odds* (traduzido, em português, como chance), que é a razão entre as probabilidades de sucesso ($P(Y = 1)$) e de fracasso ($P(Y = 0)$), ou seja,

$$\textit{odds} = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1 - P(Y=1)}.$$

Exemplo 2: Modelo de Regressão Logística

Um modelo de Regressão Logística é expresso como

$$\ln\left(\frac{P(Y_i = 1|\boldsymbol{x}_i)}{1 - P(Y_i = 1|\boldsymbol{x}_i)}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip},$$

em que x_{i1}, x_{i2}, ..., x_{ip} *correspondem à i*-ésima observação das *p* covariáveis e β_0, β_1, ..., β_p são os parâmetros do modelo.

Exemplo 2: Modelo de Regressão Logística

Equivalentemente, o modelo de Regressão Logística é expresso como

$$P(Y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}{1+\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})}.$$

Modelos lineares generalizados mistos (GLMM)

O modelo GLMM também pode ser especificado em dois estágios:

- No primeiro, assumimos que a distribuição condicional da resposta Y_{ij} dado os efeitos aleatórios $\boldsymbol{b}_i$ pertence à família exponencial linear. Então, consideramos

$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^{\top} \boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\top} \boldsymbol{b}_i \,,$$

em que g é uma função de ligação e $\boldsymbol{x}_{ij}^{\top}$ e $\boldsymbol{z}_{ij}^{\top}$ denotam, respectivamente, a j-ésima linha das matrizes $\mathbf{X}_i$ e $\mathbf{Z}_i$, correspondentes à especificação dos efeitos fixos e aleatórios, respectivamente;

- No segundo estágio, usualmente, supomos que $\boldsymbol{b}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{G}(\boldsymbol{\theta}))$ embora, em teoria, outras distribuições possam ser consideradas.

Observação

O modelo linear misto é um caso particular do modelo GLMM se a função de ligação g for a função identidade. Neste caso,

$$\mu_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{z}_{ij}^{\mathrm{T}} \boldsymbol{b}_i.$$

Observações

- Modelos lineares generalizados mistos têm uma estrutura similar àquela considerada para os modelos lineares mistos, que, na realidade, constituem um caso particular daqueles;

- No entanto, a relação entre os parâmetros de localização e de covariância por eles induzida traz mais dificuldades analíticas e de interpretação.

Exemplo 3

- Um experimento foi realizado na Faculdade de Medicina da USP com o objetivo de comparar pacientes com hepatite C tratados com dois inibidores de protease (*Telaprevir* e *Boceprevir*) relativamente à ocorrência de disfunção renal durante o período de observação de 48 semanas;

- A resposta (ocorrência de disfunção renal) assume o valor 1 quando o nível de creatinina numa determinada semana tem valor 15% maior que o correspondente valor basal;

- Um dos objetivos foi comparar os dois tratamentos com relação às chances de ocorrência de disfunção renal ao longo do período de observação controlando outras covariáveis (sexo, idade, hipertensão arterial etc.).

Figura 1: Primeiras observações do conjunto de dados, destacando-se as variáveis usadas para obter os valores da variável resposta.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	tratam	indiv	sem	sexo	idade	HAS	DM	cirrose	anem2sem	creatinina	creatmais15	cgault	cgaultmenos15	mdrd	mdrdmenos15
2	TVR	1	0	F	55	1	0	1	1	0,55	0,63	135,20	114,92	138,85	118,03
3	TVR	1	1	F	55	1	0	1	1	0,85	0,63	82,28	114,92	84,02	118,03
4	TVR	1	2	F	55	1	0	1	1	0,76	0,63	95,07	114,92	95,60	118,03
5	TVR	1	3	F	55	1	0	1	1	0,83	0,63	88,50	114,92	86,36	118,03
6	TVR	1	6	F	55	1	0	1	1	0,91	0,63	80,17	114,92	77,66	118,03
7	TVR	1	7	F	55	1	0	1	1	0,81	0,63	89,20	114,92	88,83	118,03
8	TVR	1	9	F	55	1	0	1	1	0,82	0,63	90,80	114,92	87,58	118,03
9	TVR	1	10	F	55	1	0	1	1	0,67	0,63	111,43	114,92	110,57	118,03
10	TVR	1	11	F	55	1	0	1	1	0,72	0,63	100,77	114,92	101,76	118,03
11	TVR	1	12	F	55	1	0	1	1	0,85	0,63	81,46	114,92	84,02	118,03
12	TVR	1	20	F	55	1	0	1	1	0,89	0,63	76,67	114,92	79,68	118,03
13	TVR	1	27	F	55	1	0	1	1	0,77	0,63	85,23	114,92	94,17	118,03
14	TVR	1	31	F	55	1	0	1	1	0,80	0,63	81,28	114,92	90,11	118,03

Fonte: Elaboração própria.

Exemplo 3

Para efeito didático, suponha que a ocorrência de disfunção renal dependa apenas do tempo de tratamento e que cada indivíduo tenha uma susceptibilidade própria.

Exemplo 3

Considerando que a distribuição condicional da ocorrência de disfunção renal para o i-ésimo indivíduo no instante t_{ij} possui distribuição Bernoulli, um GLMM é

$$\ln\left(\frac{P(Y_{ij} = 1|t_{ij}, b_i)}{1 - P(Y_{ij} = 1|t_{ij}, b_i)}\right) = \beta_0 + b_i + \beta_1 t_{ij},$$

em que $Y_{ij} = 1$ corresponde à ocorrência de disfunção renal para o i-ésimo indivíduo na semana t_{ij}, b_i é um efeito aleatório com distribuição $N(0, \sigma^2)$, $\beta_0 + b_i$ representa o logaritmo da *odds* (log-*odds*) de ocorrência de disfunção renal para o i-ésimo indivíduo no início do estudo ($t_{ij} = 0$) e β_1 indica o correspondente logaritmo da razão de *odds* para duas semanas consecutivas.

Exemplo 3

Equivalentemente,

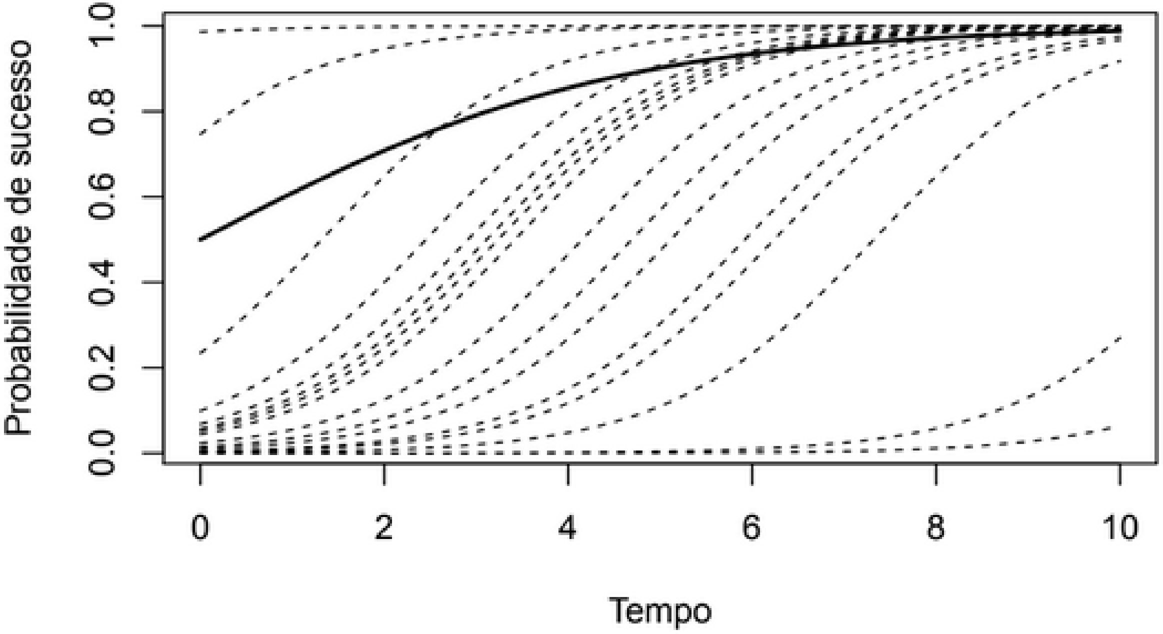
$$P(Y_{ij} = 1|t_{ij}, b_i) = \frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1+\exp(\beta_0 + b_i + \beta_1 t_{ij})}.$$

Exemplo 3

- O fato de o parâmetro β_1 estar associado à mudança no logaritmo da *odds* de ocorrência de disfunção renal para um indivíduo específico sugere que modelos lineares generalizados mistos não são apropriados para situações em que o interesse recai no parâmetro populacional correspondente;

- Esse parâmetro marginal é o valor esperado (em relação à distribuição dos efeitos aleatórios) dos parâmetros individuais (β_1).

Figura 2: Gráfico com as probabilidade de sucesso para um exemplo hipotético.



Fonte: https://www.ime.usp.br/~jmsinger/MAE0610/Singer&Nobre&Rocha2018jun.pdf

Exemplo 3

Na Figura 2, nota-se que a resposta marginal não tem o mesmo padrão que as respostas individuais.

Modelos baseados em Equações de Estimação Generalizadas (GEE)

Modelos baseados em GEE focam diretamente a distribuição marginal da variável resposta, sem referência aos efeitos aleatórios, com a especificação apenas do seu valor esperado e de sua variância

$$E(Y_{ij}) = \mu_{ij} \text{ e } Var(Y_{ij}) = \phi^{-1}\, \nu(\mu_{ij}),$$

em que $\nu(\mu_{ij})$ é uma função do valor esperado e ϕ é um parâmetro de escala.

Modelos baseados em Equações de Estimação Generalizadas (GEE)

- 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 variável resposta;

- Não é necessário ser balanceado;

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

Modelos baseados em Equações de Estimação Generalizadas (GEE)

A relação entre a resposta esperada e as variáveis exploratórias é especificada como

$$g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_p x_{ijp},$$

em que g é uma função de ligação, possivelmente não linear.

Modelos baseados em Equações de Estimação Generalizadas (GEE)

Os estimadores dos parâmetros $\boldsymbol{\beta}$ são obtidos por meio das Equações de Estimação Generalizadas

$$\sum_{i=1}^{n} D_i^{\mathrm{T}} V_i^{-1} (Y_i - \mu_i) = 0,$$

em que $\boldsymbol{V}_{\mathrm{i}} = \mathrm{Var}(\boldsymbol{Y}_{\mathrm{i}})$, $\boldsymbol{D}_i = \frac{\partial \mu i}{\partial \beta}$ e $\mu_i = g^{-1}(\boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{\beta})$.

Modelos baseados em Equações de Estimação Generalizadas (GEE)

Com a finalidade de incorporar a estrutura de covariância intraunidades amostrais, considera-se uma matriz de covariâncias de trabalho (*working covariance matrix*) definida como

$$\text{Var}(\boldsymbol{Y}_i) = \boldsymbol{V}_i = \phi\, \boldsymbol{A}_i^{1/2}\, \boldsymbol{R}_i(\boldsymbol{\theta})\, \boldsymbol{A}_i^{1/2},$$

em que A_i é uma matriz diagonal formada por $\text{Var}(Y_{ij})$, $\boldsymbol{R}_i(\boldsymbol{\theta})$ é matriz de correlação de trabalho e ϕ é um parâmetro de dispersão/escala.

Observações

- Os estimadores dos parâmetros de regressão são obtidos como solução para as equações de estimação generalizadas;

- Se $\boldsymbol{R}_i(\boldsymbol{\theta})$ for a verdadeira matriz de correlações intraunidades amostrais de Y_i, então $\boldsymbol{R}_i(\boldsymbol{\theta}) = \mathrm{Var}(Y_i)$.

- Para dados longitudinais não correlacionados, a matriz de covariâncias de trabalho considera a estrutura de independência, ou seja, $\boldsymbol{R}_i(\theta) = I_n$.

Exemplo 4: 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);

- Variável resposta Binária: 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.

Tabela 1: Estimativa dos parâmetros do modelo GEE.

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	

Exemplo 4: Mecanismo Evacuatório de Récem-Nascidos

Interpretações:

- **Idade:** A razão de *odds* é $\exp(-0, 124)$ = 0,83 (0,81;0,95), isto significa que, com o aumento de um ano na idade, a *odds* de dificuldade em evacuar reduz em 17%;

- **Sexo:** A razão de *odds* *é* $\exp(-0, 485)$ = 0,615 (0,40;0,947), isto significa que a *odds* de meninos terem dificuldade em evacuar é 39% menor do a *odds* das meninas;

- **Aleitamento:** A razão de *odds é* $\exp(0, 796)$ = 2,22 (1,23;3,99), isto significa que a *odds* de recem nascidos que tiveram aleitamento artificial ter dificuldade em evacuar é 2,2 vezes a *odds* daqueles que tiveram aleitamento no peito.

Observações

Utilizando-se um modelo linear generalizado misto:

- 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.

Observações

Utilizando-se um modelo de regressão logística misto:

- De forma a termos a interpretação usual de razão de *odds*, RC = e^{β}, para o aumento em uma unidade de x, devemos cancelar o efeito aleatório b_i;

- Assim, β é o log-*odds* da resposta por aumento de uma unidade em x, para qualquer indivíduo tendo uma propensão de resposta positiva b_i;

- 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.

Exemplo 5: 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);

- Variável resposta binária: 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.

Tabela 2: Estimativa dos parâmetros do modelo de regressão logística misto.

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.

Especialização em *Data Science* e Estatística Aplicada

Módulo IV – Análise de Dados Longitudinais

Prof. Dr. Márcio Luis Lanfredi Viola
lanfredi@ufscar.br

IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
FEN FACULDADE DE ENFERMAGEM
UFG UNIVERSIDADE FEDERAL DE GOIÁS