

# **CE225 - Modelos Lineares Generalizados**

Cesar Augusto Taconeli

11 de julho, 2018

## **Aula 11 - Regressão para dados binários**

# Introdução

- Interesse em modelar fenômenos aleatórios com dois desfechos possíveis (*sucesso* ou *fracasso*) como função de covariáveis;
- A distribuição binomial (e, como caso particular, a distribuição Bernoulli) surge como principal alternativa para a modelagem de dados binários;
- Grande quantidade e variedade de potenciais aplicações.

# Exemplos de motivação

- Prognóstico clínico de pacientes (ex: cura ou não cura) em função de variáveis clínicas, genéticas, comportamentais. . .
- Risco de crédito (pagamento ou não) por clientes de um banco em função de variáveis sócio-econômicas, referentes à modalidade de empréstimo, ao relacionamento do cliente com o banco. . .
- Resultado de partidas de basquete (vitória do mandante ou do visitante) em função do desempenho das equipes no campeonato, histórico de confrontos, circunstâncias da partida, . . .
- Presença ou não de certa espécie vegetal em regiões de uma floresta em função de variáveis ambientais e climáticas.

# Distribuição de Bernoulli

- A distribuição de Bernoulli associa valores 0 e 1 a cada um dos dois desfechos.
- Uma variável aleatória  $y$  tem distribuição de Bernoulli com parâmetro  $\pi$  se sua função de probabilidades é dada por:

$$f(y; \pi) = \pi^y(1 - \pi)^{1-y}, \quad y = 0, 1; \quad 0 < \pi < 1, \quad (1)$$

ou seja,  $f(0; \pi) = P(y = 0|\pi) = 1 - \pi$  e  $f(1; \pi) = P(y = 1|\pi) = \pi$ .

- Propriedades da distribuição Bernoulli:

$$E(y) = \mu = \pi; \quad \text{Var}(Y) = \mu(1 - \mu). \quad (2)$$

- A distribuição Bernoulli pertence à família exponencial com  $V(\mu) = \mu(1 - \mu)$  e  $\phi = 1$ .

# Distribuição binomial

- Considere  $n$  realizações independentes de um experimento de Bernoulli, todos com mesma probabilidade de sucesso  $\pi$ .
- Seja  $Y$  a fração de sucessos observada nas  $n$  realizações. Então:

$$f(y; n, \pi) = \binom{n}{ny} \pi^{ny} (1 - \pi)^{n-(ny)}; y = 0, \frac{1}{n}, \frac{2}{n}, \dots, 1; 0 < \pi < 1. \quad (3)$$

- A Bernoulli é um caso particular da distribuição binomial (quando  $n = 1$ ).

# Distribuição binomial

- A média e a variância de  $y$  são dadas por:

$$E(y) = \mu = \pi; \quad \text{Var}(Y) = \frac{\mu(1 - \mu)}{n}. \quad (4)$$

- A distribuição binomial pertence à família exponencial com  $V(\mu) = \mu(1 - \mu)$  e  $\phi = 1$ .

# Regressão para dados binários

- No contexto de modelos lineares generalizados, considere  $y_1, y_2, \dots, y_n$  variáveis aleatórias independentes com

$$y_i \sim \text{binomial}(m_i, \pi_i), \quad i = 1, 2, \dots, n. \quad (5)$$

- Adicionalmente, sejam  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  os vetores de covariáveis associados a cada observação;
- Especificação do modelo linear generalizado:

$$y_i | \mathbf{x}_i \sim \text{binomial}(m_i, \pi_i);$$
$$g(\pi_i) = \eta_i = \mathbf{x}_i' \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \quad (6)$$



# Escolha da função de ligação

- Dentre os requisitos para a escolha de uma função de ligação adequada, destacamos:
  - A função de ligação deve ser contínua, diferenciável e monótona;
  - Capaz de ‘mapear’ os valores de  $\pi$  no intervalo (0,1);
  - Capaz de linearizar a relação entre os componentes aleatório e sistemático do modelo;
  - Que proporcione interpretações simples.

# Escolha da função de ligação

- Boa parte dos requisitos indicados são atendidos se adotarmos, como função de ligação, a inversa de  $F$ , a função distribuição acumulada (fda) de alguma variável aleatória contínua:

$$g(\pi_i) = F^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad (7)$$

ou, de forma equivalente,

$$\pi_i = F(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}). \quad (8)$$

- Três funções usuais para MLGs com dados binários são as ligações **logito**, **probit** e **complemento log-log**, discutidas na sequência.

# Função de ligação logito

- A função de ligação logito baseia-se na fda da distribuição logística em sua forma padrão ( $\mu = 0$  e  $\sigma = 1$ ):

$$F(z) = \frac{e^z}{(1 + e^z)}. \quad (9)$$

- Assim como a distribuição normal padrão, a distribuição logística padrão é definida em todo o conjunto dos reais, com forma de sino centrada em zero.
- O MLG baseado na função de ligação logito fica definido por:

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}}}, \quad (10)$$

ou, na escala do preditor:

$$\ln \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}. \quad (11)$$

# Função de ligação probito

- A função de ligação probito fica definida pela (inversa) da fda da distribuição normal padrão, definindo o seguinte MLG:

$$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad (12)$$

ou, na escala da resposta:

$$\pi_i = \Phi(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}), \quad (13)$$

em que  $\Phi(\cdot)$  denota a fda da distribuição normal padrão.

- Na prática, as funções de ligação probito e logito têm comportamentos bastante semelhantes, sobretudo no intervalo (0.1,0.9).

# Função de ligação complemento log-log

- A função de ligação complemento log-log baseia-se na (inversa) da fda da distribuição Gumbel, definindo o seguinte MLG:

$$\ln[-\ln(1 - \pi_i)] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad (14)$$

ou, na escala da resposta,

$$\pi_i = 1 - \exp[-\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip})]. \quad (15)$$

- Diferentemente das funções de ligação probito e logito, a função de ligação complemento log-log é assimétrica em relação a  $\pi = 0.5$ , o que pode ser conveniente em algumas aplicações.

- A família de ligações Aranda-Ordaz é definida por:

$$g(\pi_i) = \ln \left[ \frac{(1 - \pi_i)^{-\alpha} - 1}{\alpha} \right], \quad (16)$$

sendo  $\alpha$  um parâmetro que pode ser estimado.

- Como caso particular, para  $\alpha = 1$  temos a função de ligação logito;
- Para  $\alpha \rightarrow 0$  temos a função de ligação complemento log-log.

# Regressão logística

- O modelo de regressão logística fica definido pelo uso da ligação logito em um MLG binomial.
- Neste caso, considerando  $y_i | \mathbf{x}_i \sim \text{binomial}(m_i, \pi_i)$ ,  $i = 1, 2, \dots, n$ , independentes, então o modelo de regressão logística fica dado por:

$$g(\pi) = \ln \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}. \quad (17)$$

- De forma equivalente, na escala da *odds*:

$$\frac{\pi_i}{1 - \pi_i} = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}. \quad (18)$$

- Na escala da probabilidade (resposta):

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}}}{e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}} + 1}. \quad (19)$$



# Regressão logística - interpretação dos parâmetros

- A interpretação dos parâmetros em um modelo de regressão logística baseia-se em razões de chances (*odds ratios*).
- Começamos pelo caso da regressão logística simples, com apenas uma covariável ( $x$ ). Assumindo que  $x$  seja uma variável numérica, então:

$$OR\{x+1, x\} = \frac{odds\{x+1\}}{odds\{x\}} = \frac{\pi_{x+1}/(1-\pi_{x+1})}{\pi_x/(1-\pi_x)} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1x}} = e^{\beta_1}, \quad (20)$$

em que  $\pi_x = P(Y = 1|x)$ .

- Assim,  $e^{\beta_1}$  corresponde ao acréscimo na chance de resposta ( $y = 1$ ) para um aumento unitário em  $x$ .

# Regressão logística - interpretação dos parâmetros

- Para um aumento de  $k$  unidades em  $x$ , verifica-se facilmente que a chance de resposta fica aumentada em  $e^{k\beta_1}$ .
- Para o caso de uma covariável dicotômica (com categorias A e B), inserida ao modelo por meio de uma variável indicadora de B, temos:

$$OR\{B, A\} = \frac{odds\{B\}}{odds\{A\}} = \frac{\pi_B/(1 - \pi_B)}{\pi_A/(1 - \pi_A)} = \frac{e^{\beta_0 + \beta_1 \times 1}}{e^{\beta_0 + \beta_1 \times 0}} = e^{\beta_1}, \quad (21)$$

em que  $\pi_A$  é a probabilidade de resposta para um indivíduo da categoria A.

- Assim,  $e^{\beta_1}$  corresponde à razão das chances de resposta para as categorias B e A.

# Regressão logística - interpretação dos parâmetros

- Se houvesse uma terceira categoria (C), então seriam necessárias duas variáveis indicadoras ( $x_1$ , indicadora de B;  $x_2$ , indicadora de C). Assim, teríamos:

$$\ln \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \quad (22)$$

- As razões de chances ficariam dadas por:

$$OR\{B, A\} = \frac{\text{odds}\{B\}}{\text{odds}\{A\}} = \frac{e^{\beta_0 + \beta_1 \times 1 + \beta_2 \times 0}}{e^{\beta_0 + \beta_1 \times 0 + \beta_2 \times 0}} = e^{\beta_1}; \quad (23)$$

$$OR\{C, A\} = \frac{\text{odds}\{C\}}{\text{odds}\{A\}} = \frac{e^{\beta_0 + \beta_1 \times 0 + \beta_2 \times 1}}{e^{\beta_0 + \beta_1 \times 0 + \beta_2 \times 0}} = e^{\beta_2}; \quad (24)$$

$$OR\{C, B\} = \frac{\text{odds}\{C\}}{\text{odds}\{B\}} = \frac{e^{\beta_0 + \beta_1 \times 0 + \beta_2 \times 1}}{e^{\beta_0 + \beta_1 \times 1 + \beta_2 \times 0}} = e^{\beta_2 - \beta_1}. \quad (25)$$

# Regressão logística - interpretação dos parâmetros

- Caso o preditor linear contenha múltiplas variáveis, as interpretações são idênticas, devendo-se ressaltar, no entanto, que a interpretação da razão de chances calculada para uma particular variável, é válida fixando os valores das demais variáveis.
- **Aplicação:** Pesquisa de opinião pública, plebiscito no Chile. **Vamos ao R!**