Modelos Lineares Generalizados

Emilly Malveira de Lima

Análise de Dados Categóricos
Universidade Federal de Minas Gerais - UFMG

Modelo de regressão linear

Regressão linear

Tem o interesse de modelar o comportamento médio de uma variável de interesse por meio da combinação de covariáveis explicativas.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + \varepsilon_i, \quad i = 1, ..., n.$$

$$E[Y_i|X_i] = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} = \underbrace{\mathbf{X}_i'\boldsymbol{\beta}}_{\text{Preditor linear}}$$

- $\varepsilon_i's$ são independentes e têm a mesma distribuição, $N(0, \sigma^2)$
- $Y_i \sim N(E[Y_i|X_i], \sigma^2)$
- Suposições do modelo: linearidade, independência, homocedasticidade e **normalidade**

Modelos lineares generalizados

- Os MLGs foram propostos por Nelder and Wedderburn (1972) como uma extensão do modelo linear normal;
- são modelos para análise de dados em que a suposição de normalidade não é plausível;
- é possível modelar variáveis de interesse que assumem a forma de **contagem**, contínuas simétricas e assimétricas, **binárias** e **categóricas**;
- assume-se que a densidade da função de distribuição da variável resposta pertence à família exponencial.
- a modelagem da média não é feita de forma direta. É necessário escolher uma **função** que é capaz de relacionar as covariáveis com a média da variável resposta.

Estrutura do MLG

Há 3 perguntas a serem feitas na hora da construção de um MLG

1. Qual é a distribuição dos dados?
   O vetor $\boldsymbol{Y}$ geralmente consiste em uma amostra aleatória proveniente de uma distribuição de probabilidade com densidade na família exponencial.

2. Quais são os preditores? Variáveis explicativas que possivelmente estão associadas à variável resposta.

3. Qual é a função da média que será modelada como função linear dos preditores? *Função de ligação*

Estrutura do MLG

Há 3 perguntas a serem feitas na hora da construção de um MLG

1. Qual é a distribuição dos dados?
   O vetor $\boldsymbol{Y}$ geralmente consiste em uma amostra aleatória proveniente de uma distribuição de probabilidade com densidade na família exponencial.

2. Quais são os preditores? Variáveis explicativas que possivelmente estão associadas à variável resposta.

3. Qual é a função da média que será modelada como função linear dos preditores? *Função de ligação*

Estrutura do MLG

Há 3 perguntas a serem feitas na hora da construção de um MLG

1. Qual é a distribuição dos dados?
   O vetor $\boldsymbol{Y}$ geralmente consiste em uma amostra aleatória proveniente de uma distribuição de probabilidade com densidade na família exponencial.

2. Quais são os preditores? Variáveis explicativas que possivelmente estão associadas à variável resposta.

3. Qual é a função da média que será modelada como função linear dos preditores? *Função de ligação*

Estrutura do MLG

Há 3 perguntas a serem feitas na hora da construção de um MLG

1. Qual é a distribuição dos dados?
   O vetor $\boldsymbol{Y}$ geralmente consiste em uma amostra aleatória proveniente de uma distribuição de probabilidade com densidade na família exponencial.

2. Quais são os preditores? Variáveis explicativas que possivelmente estão associadas à variável resposta.

3. Qual é a função da média que será modelada como função linear dos preditores? *Função de ligação*

Família exponencial

De maneira geral, uma distribuição é dita ser da família exponencial se sua densidade pode ser escrita na seguinte forma:

Estrutura da família exponencial

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

- y é a variável de interesse
- θ é parâmetro natural (*função de ligação canônica*)
- ϕ parâmetro de escala
- $a, b,$ e c são funções específicas que derterminam unicamente a distribuição.

Propriedades:

$\mu = E(Y) = b'(\theta) = \frac{db(\theta)}{d\theta}$ e

$\sigma^2 = Var(Y) = b''(\theta)a(\phi) = \frac{d^2 b(\theta)}{d\theta^2}a(\phi)$

Tipos de respostas e distribuições

- **Contagem**: modelo de Poisson
- **Binária**: modelo Bernoulli
- **Categórica:** modelo multinomial
- **Contínua assimétrica:** modelo gama
- **Outras:** etc

Resposta Poisson

Contagem/Taxas de ocorrência de eventos

- Número de acidentes de carro;
- Número de mortes por uma determinada doença;
- Número de casos de dengue em uma região (município, estado...)

Poisson

Função de probabilidade

$$f_Y(y|\lambda) = \frac{\lambda^y}{y!}e^{-\lambda},\ y = 0, 1, 2, \ldots\ ;\ \lambda > 0.$$

Forma da família exponencial

$$f_Y(y|\lambda) = exp\{ylog(\lambda) - \lambda + [-log(y!)]\}$$

$$\theta = log(\lambda);\ b(\theta) = \lambda = exp(\theta); a(\phi) = 1;\ c(y, \phi) = -log(y!)$$

Média e variância

$$\mu = b'(\theta) = exp(\theta) = \lambda$$

$$\sigma^2 = b''(\theta)a(\phi) = \lambda$$

Resposta Binária

Binária Quando temos o intuito de modelar uma variável resposta 0-1 ("fracasso-sucesso"), ou seja, a probabilidade de sucesso de algum evento de interesse

- probabilidade de um paciente desenvolver algum tipo de doença;
- probabilidade de um indivíduos possuir plano de saúde;
- probabilidade de um cliente de banco ser inadimplente;

Bernoulli

Função de probabilidade

$$f_Y(y|\pi) = \pi^y(1-\pi)^{1-y},\ y = \{0,1\},\ 0 < \pi < 1.$$

Forma da família exponencial

$$f_Y(y|\pi) = exp\left\{ylog\left(\frac{\pi}{1-\pi}\right) - [-log(1-\pi)]\right\}$$

$$\theta = log\left(\frac{\pi}{1-\pi}\right);\ b(\theta) = log(1+exp(\theta)); a(\phi) = 1;\ c(y,\phi) = 0$$

Média e variância

$$\mu = b'(\theta) = \frac{exp(\theta)}{1+exp(\theta)} = \pi$$

$$\sigma^2 = b''(\theta)a(\phi) = \frac{exp(\theta)}{[1+exp(\theta)]^2} = \pi(1-\pi)$$

Função de ligação

- Sejam $X_1, X_2, \ldots, X_p$ um conjunto de covariáveis.
- A média da variável resposta

$$\mu = E(Y)$$

é relacionada com $X_1, X_2, \ldots, X_p$ através de uma função de ligação g

- A relação entre μ e η é definida por

$$g(\mu) = g(E[Y]) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = \mathbf{X}'\boldsymbol{\beta}$$

- geralmente são utilizadas as funções canônicas, ou seja, $\theta = \eta$

Função de ligação

Exemplo: Poisson (Regressão de Poisson)

Sob a ligação canônica devemos ter

$$\theta = log(\lambda) = \eta = \mathbf{X}'\boldsymbol{\beta}$$

então

$$g(E(Y)) = g(\lambda) = log(\lambda)$$

Para o modelo de Poisson g é chamada de ligação *log*

Função de ligação

Exemplo: Bernoulli (Regressão logística)

Sob a função de ligação canônica temos

$$\theta = log\left(\frac{\pi}{1-\pi}\right) = \eta = \mathbf{X}'\beta$$

então

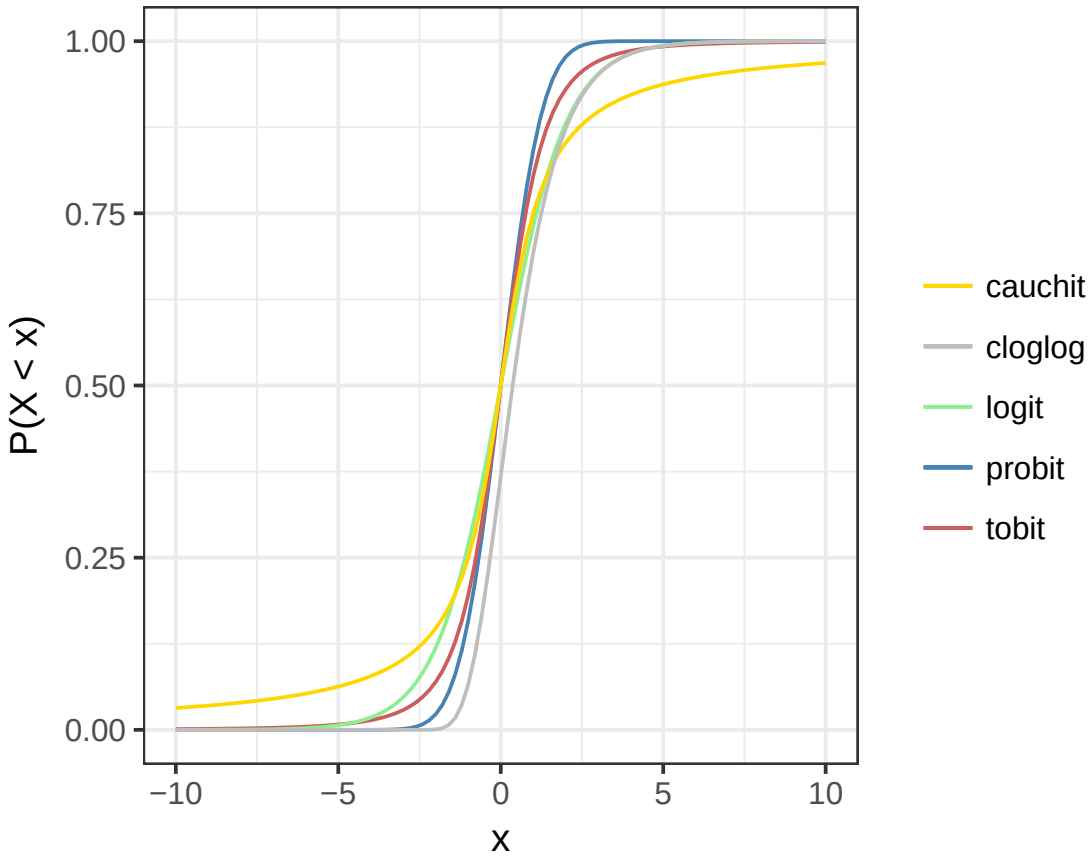
$$g(E(Y)) = g(\pi) = log\left(\frac{\pi}{1-\pi}\right)$$

Para o modelo de Bernoulli g é chamada de ligação *logit*.

- O modelo de regressão logística é expresso por

$$p(y|\mathbf{x}) = \frac{exp(\mathbf{X}'\beta)}{1 + exp(\mathbf{X}'\beta)}$$

Outras ligações na regressão logística



- Mais utilizadas: logit, probit e complemento log-log.

Inferência nos MLG's

- Inferência baseada na teoria assintótica de máxima verossimilhança;
- Métodos numéricos: Newton Raphson e Escore de Fisher

Exemplo: Função de verossimilhança da Bernoulli

$$L(\beta) = \prod_{i=1}^{n} f_Y(y_i|\pi) = \pi^{y_i}(1-\pi)^{1-y_i} = \prod_{i=1}^{n} \frac{exp(y_i \mathbf{X}_i' \beta)}{1 + exp(\mathbf{X}_i' \beta)}$$

Função de log-verossimilhança

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \{y_i \mathbf{X}_i' \beta - log[1 - exp(\mathbf{X}_i' \beta)]\}$$

Porém, nos MLG's os estimadores dos $\beta's$ **NÃO** são encontrados de forma analítica.

Newton-Raphson

- o algoritmo de *Newton-Raphson* é comumente utilizado para calcular o EMV de β.

Algoritmo de Newton-Raphson

Seja $l'(\beta)$ o logaritmo da função de verossimilhança. A cada passo do algoritmo NR a estimativa β^t é atualizada por

$$\beta^{t+1} = \beta^t + [l''(\beta)]^{-1} l'(\beta^t)$$

em que

$l'(\beta) = \left(\frac{\partial l(\beta)}{\partial \beta_0}, \frac{\partial l(\beta)}{\partial \beta_1}, \ldots, \frac{\partial l(\beta)}{\partial \beta_p} \right)$ conhecido com vetor *escore*

$l''(\beta)$ é a matriz de segundas derivadas com $(p+1)x(p+1)$ elementos

Inferência nos GLM's

Teste de Hipóteses

- Baseados na teoria assintótica de MV;

- usualmente três estatísticas específicas para realizar testes sobre β's:
  - i Wald
  - ii Razão de verossimilhanças
  - iii Escore

1) Teste de Wald parar testar a hipótese de nulidade do efeito das covariáveis, ou seja, $H_0 : \beta_j = 0$.

- A estatística de teste $X_j^2 = \left(\frac{\hat{\beta}_j}{EP(\hat{\beta}_j)}\right)^2$ segue uma distribuição χ^2 com 1 gl.

2) TRV para testar modelos aninhados.

$$TRV = -2\left[\frac{\ell(H_0)}{\ell(H_1)}\right] \underset{\sim}{\text{assintótica}} \chi_{p-q}$$

Sendo p o nº de parâmetros do modelo sem restrição, e q do modelo com restrição.

No R

glm()

- **formula**: $y \sim x1 + x2 + ...$
- **family**: binomial(link = "logit")
  poisson(link = "log")
  gaussian(link = "identity")
  Gamma(link = "inverse")
- **data**
- **offset**

TRV

anova(**modelo sob restrição**, **modelo sem restição**)

Referências

Nelder, J., & Wedderburn, R. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A (General), 135(3), 370-384.

Cordeiro, G. M., & Demétrio, C. G. (2008). Modelos lineares generalizados e extensões. Sao Paulo.

Dobson, A. J., & Barnett, A. (2008). An introduction to generalized linear models. CRC press.