Profª. Msc. Viviane Costa Silva

Disciplina de Seminários - UEPB - Universidade Estadual da Paraíba
Profª Drª Ana Patricia Bastos Peixoto

MODELOS LINEARES GENERALIZADOS (MLG) EM PYTHON: DA TEORIA À PRÁTICA

OBJETIVOS DE APRENDIZAGEM

Ao final da aula, os alunos deverão ser capazes de

- Entender a estrutura matemática de um GLM: família, função de ligação e preditor linear.
- Escolher família e link apropriados para tipos de resposta (binária, contagem, contínua positiva).
- Ajustar GLMs em Python usando *statsmodels* (e mostrar alternativas como *scikit-learn* para *logistic*) e interpretar coeficientes.
- Diagnosticar problemas de ajuste (*overdispersion*, resíduos, influência).
- Ajustar variantes (NegBin, quasi-Poisson, Gamma) quando necessário.

MOTIVAÇÃO

- Quando OLS falha: binária, contagem, proporção, dados assimétricos.

O OLS assume que o erro tem distribuição Normal (Gaussiana) e que a variância é constante (homocedasticidade). Quando a variável resposta *Y* tem características específicas (como as listadas abaixo), essas premissas são violadas, levando a estimativas ineficientes ou viesadas.

Tipo de Dado	OLS Falha Porque...	Exemplo Curto	Solução Padrão (GLM)
Binário (0 ou 1)	Viola a **Normalidade** e a **Homocedasticidade**. Previsões fora de $[0, 1]$.	Taxa de sucesso (0/1)	**Regressão Logística** (Distribuição Binomial, Link Logit)
Contagem ($0, 1, 2, \ldots$)	Viola a **Normalidade** e a **Variância** (Variança $\approx$ Média). Previsões < 0.	Número de acidentes por mês	**Regressão de Poisson** (Distribuição de Poisson, Link Log)
Proporção ($\in [0, 1]$)	Se a proporção é calculada por $(0, 1)$, o OLS ainda sofre com a restrição do intervalo.	Proporção de votos em um candidato	**Regressão Beta** ou **Binomial** (Link Logit/Probit)
Dados Assimétricos (Ex: Renda)	A distribuição da resposta é muito enviesada (cauda longa), violando a Normalidade.	Distribuições com cauda direita longa (Ex: dados de custo, tempo de espera)	**Regressão Gama** (Distribuição Gama, Link Log/Inversa)

ESTRUTURA GERAL DE UM MLG

$$g(E[Yi]) = \eta i = Xi\beta$$

em que,

Yi : variável resposta

$g(\cdot)$: função de ligação

Xi : vetor de preditores

β : parâmetros do modelo

FAMÍLIA EXPONENCIAL

Os Modelos Lineares Generalizados (GLMs) foram criados para unificar vários tipos de regressão (linear, logística, Poisson etc.) sob uma única estrutura matemática. Isso é possível porque todas essas distribuições pertencem a um mesmo grupo: a família exponencial de distribuições.

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

onde:

Símbolo	Significado
θ	**Parâmetro canônico** (relacionado à média)
ϕ	**Parâmetro de dispersão** (variança)
$a(\phi)$, $b(\theta)$, $c(y,\phi)$	Funções que definem cada distribuição específica

FUNÇÃO DE LIGAÇÃO (*LINK FUNCTION*)

A função de ligação (link) conecta a média da variável resposta $\mu i = E(Yi)$ ao componente linear (a combinação dos preditores)

$$g(\mu i) = \beta 0 + \beta 1 X1i + \beta 2 X2i + \cdots + \beta p Xpi$$

A função de ligação é a ponte que transforma a média do fenômeno em algo que possa ser modelado como uma combinação linear dos preditores.

COMPONENTE LINEAR

O componente linear é a parte “regressão” do GLM — a combinação ponderada dos preditores por seus coeficientes.

$$\eta i = \beta 0 + \beta 1 X 1 i + \beta 2 X 2 i + \cdots + \beta p X p i$$

onde:

- ηi é o previsor linear (linear predictor),
- X_{ji} são as variáveis explicativas,
- βj são os coeficientes estimados.

FAMÍLIAS E LINKS

Família (family)	Distribuição da variável resposta	Link padrão (link)	Outros links disponíveis	Exemplo de aplicação
Gaussian	Normal (contínua)	identity	log, inverse_power	Regressão linear clássica (ex: prever altura, preço, temperatura)
Binomial	Bernoulli (0/1) ou número de sucessos em n ensaios	logit	probit, cloglog, log, cauchit	Regressão logística (ex: sobrevivência, aprovação, classificação binária)
Poisson	Contagem (inteiros ≥ 0)	log	identity, sqrt	Contagem de eventos (ex: número de chamadas, casos de doença, focos de incêndio)
Gamma	Contínua, positiva e assimétrica	inverse	identity, log	Modelos de duração ou custo (ex: tempo até falha, custo médico)
Inverse Gaussian	Contínua, positiva (com cauda pesada)	inverse_squared	identity, inverse, log	Tempos de sobrevivência e processos físicos
Negative Binomial	Contagem com sobredispersão (var > média)	log	identity, sqrt	Contagens com variabilidade extra (ex: casos de dengue, sinistros)
Tweedie	Contínua e discreta (mista)	log	—	Modelos mistos: dados com zero-inflados e contínuos (ex: prêmios de seguro)

Profª. Msc. Viviane Costa Silva - UFLA

INTERPRETAÇÃO DOS COEFICIENTES EM GLM

Escala do Link

Os coeficientes (β) em um GLM não estão diretamente na escala da variável resposta — eles estão na escala da função de ligação (link).
Para interpretá-los, é preciso voltar à escala original da média (μ).

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

$$\mu_i = g^{-1}(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})$$

Profª. Msc. Viviane Costa Silva - UFLA

EXEMPLO 1 — LOGIT (REGRESSÃO LOGÍSTICA)

Modelar a probabilidade de sobrevivência (0 = não, 1 = sim) no Titanic.

Modelo $$logit(p) = \beta 0 + \beta 1 Idade$$

Suponha $$\beta 1 = -0.05$$

Interpretação:

- Cada aumento de 1 ano na idade reduz o log-odds de sobrevivência em 0.05.
- Transformando para odds ratio:

$$exp(-0.05) = 0.95$$

As odds (chances) de sobrevivência diminuem 5% por ano de idade.

EXEMPLO 2 — REGRESSÃO DE CONTAGEM (POISSON)

Modelar o número de chamadas atendidas por um funcionário por dia.

Modelo $$log(\mu i) = \beta 0 + \beta 1 \cdot HorasTrabalho$$

Suponha $$\beta 1 = 0,1$$

Interpretação:

- Cada hora adicional de trabalho aumenta o log da média de chamadas em 0,1.
- Transformando para odds ratio:

$$e\hat{}0.1 = 1.105$$

A média esperada de chamadas cresce 10,5% por hora de trabalho.

EXEMPLO 3 — REGRESSÃO GAMMA (TEMPO DE ESPERA)

Modelar o tempo médio de espera em uma fila (variável contínua positiva).

Modelo $$log(\mu i) = \beta 0 + \beta 1 \cdot Funcionarios$$

Suponha $$\beta 1 = -0,3$$

Interpretação:

- Cada funcionário adicional reduz o log da média do tempo de espera em 0,3.
- Transformando para odds ratio:

$$e\hat{}0.3 = 0.74$$

O tempo médio de espera é 26% menor para cada funcionário extra.

ESTIMAÇÃO E VEROSSIMILHANÇA

Nos Modelos Lineares Generalizados (GLM), os parâmetros β são estimados pelo método da Máxima Verossimilhança (MLE) não mais pelo método dos Mínimos Quadrados (OLS) da regressão linear clássica.

- Em OLS (regressão linear comum), ajustamos a reta que minimiza os erros quadráticos.
- Em GLM, ajustamos o modelo que torna os dados observados mais prováveis sob a distribuição assumida.

Profª. Msc. Viviane Costa Silva - UFLA

ESTIMAÇÃO E VEROSSIMILHANÇA

A verossimilhança (Likelihood) mede o quão bem um conjunto de parâmetros explica os dados:

$$L(\beta) = P(y1, y2, \ldots, yn \mid \beta)$$

Por conveniência, trabalha-se com o **logaritmo da verossimilhança**:

$$\ell(\beta) = logL(\beta)$$

Profª. Msc. Viviane Costa Silva - UFLA

DIAGNÓSTICO GERAL EM GLM

Após ajustar o modelo, precisamos verificar se o GLM está descrevendo bem os dados.
Diagnósticos ajudam a identificar ajustes ruins, outliers, variabilidade excessiva e pontos influentes.

RESÍDUOS

Resíduos indicam diferenças entre o observado e o previsto.
Existem várias formas de analisá-los em GLM:

Tipo de Resíduo	**Fórmula / Ideia**	**Interpretação prática**
Pearson	$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$	Mede discrepância padronizada (espera-se ~N(0,1))
Deviance	$r_{D,i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2(d_i)}$	Mede contribuição à deviance total — útil para detectar outliers
Deviance total	Soma de $r_{D,i}^2$	Mede ajuste global do modelo

RESUMO DIDÁTICO

Regressão Linear (OLS)	GLM (MLE)
Minimiza soma dos erros quadrados	Maximiza probabilidade dos dados
Pressupõe erros normais	Usa família de distribuições (Exponencial)
Estimador: $\hat{\beta} = (X'X)^{-1}X'y$	Estimador via iteração numérica (Newton–Raphson / IRLS)

ENTRE EM CONTATO

E-mail

viviane.silva6@estudante.ufla.br

Redes sociais

@viviane_costass

Telefone

(81) 99749-3008