

**MBA
USP
ESALQ**

**Supervised Machine
Learning: Modelos
Logísticos
Binários e Multinomiais**

Prof. Dr. Luiz Paulo Fávero

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98

MODELOS LOGÍSTICOS



Fundamentação teórica e conceitos

Especificação do modelo e funções de ligação canônica

Modelos logísticos binários e multinomiais

Estimação dos parâmetros por máxima verossimilhança

Cutoff, sensibilidade, especificidade, curva ROC e índice GINI

Estimações em Python

Modelos Lineares Generalizados (GLM)

$$Y = f(X_1, X_2, X_3, \dots, X_k)$$

Modelo de Regressão	Característica da Variável Dependente	Distribuição
Linear	Quantitativa	Normal
Com Transformação de Box-Cox	Quantitativa	Normal Após a Transformação
Logística Binária	Qualitativa com 2 Categorias (<i>Dummy</i>)	Bernoulli
Logística Multinomial	Qualitativa M ($M > 2$) Categorias	Binomial
Poisson	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson
Binomial Negativo	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson-Gama





Regressão Logística Binária

- Técnica supervisionada de machine learning utilizada para explicar ou prever a probabilidade de ocorrência de determinado evento em função de uma ou mais variáveis explicativas.
- **Variável dependente:** binária.
→ Resultados interpretados em termos de probabilidades.
- **Variáveis preditoras X :** métricas ou não métricas.



Objetivos da Técnica

- **Atribuição de Probabilidades:**
Estimar a probabilidade de ocorrência de determinado evento ou de que uma observação venha a se enquadrar nessa ou naquela categoria.
- **Classificação em categorias:**
Classificar indivíduos ou observações em categorias específicas.



Conceitos:

Probabilidade

Chance (*odds*)

Logito

■ Conceito de Probabilidade

Seja Y a resposta a um estímulo (sim ou não) - pode ser a preferência por um produto, adimplência, aprovação em um curso, etc.

- p : probabilidade da resposta “sim”.
- $1 - p$: probabilidade da resposta “não”.

Conceito de Chance (*Odds*)

- Chance (*odds*) de ocorrência de um evento:

$$chance = \frac{p}{1 - p} \quad \begin{array}{l} \text{(Evento)} \\ \text{(Não Evento)} \end{array}$$

Exemplos: se $p = 0,50$; chance = 1 (1 para 1)
se $p = 0,75$; chance = 3 (3 para 1)
se $p = 0,25$; chance = $\frac{1}{3}$ (1 para 3)

Conceito de Logito

- **Logito: logaritmo natural da chance de ocorrência de uma resposta do tipo “sim”.**

E, a partir do logito, define-se a expressão da probabilidade de ocorrência do evento em estudo, em função das variáveis explicativas.



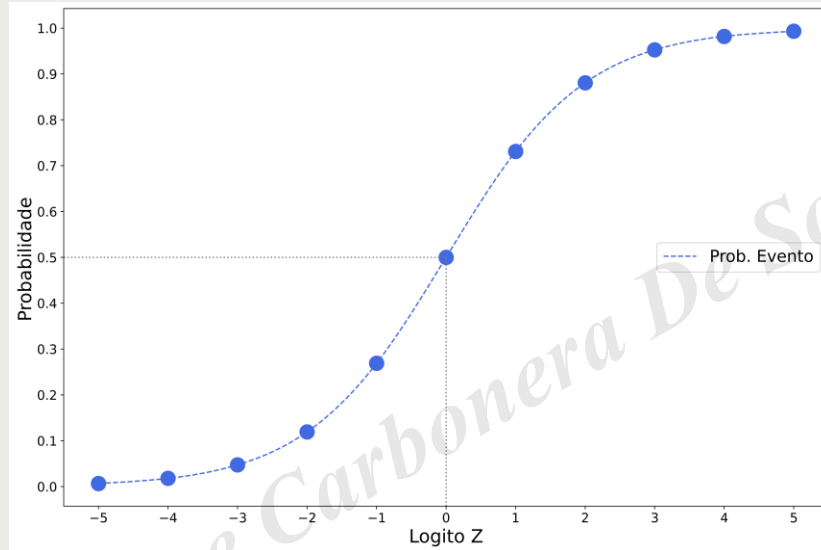
$$\text{logito} = Z = \ln \left(\frac{p}{1-p} \right)$$

$$e^{\text{logito}} = e^Z = \frac{p}{1-p} = \text{odds}$$

$$p = \frac{e^Z}{1 + e^Z} = \frac{1}{1 + e^{-Z}}$$



O Modelo de Regressão Logística Binária



A curva logística, ou sigmoide, descreve a relação entre a probabilidade associada à ocorrência de determinado evento e um conjunto de variáveis preditoras.

A função logística assume valores entre 0 e 1 para qualquer Z entre $-\infty$ e $+\infty$

$$p_i = \frac{1}{1 + e^{-Z_i}} = \frac{1}{1 + e^{-(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})}}$$

Função Logística

$$p_i = \frac{1}{1 + e^{-Z_i}} = \frac{1}{1 + e^{-(\alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki})}}$$

- Definida para que se estabeleça a probabilidade de ocorrência de determinado evento e a importância das variáveis explicativas para esta ocorrência.
- Estimação dos parâmetros:** processo iterativo para maximizar o acerto da probabilidade de ocorrência de um evento à sua real ocorrência (**Método de Máxima Verossimilhança**).
- Os **resultados** atribuíveis à variável dependente estarão entre 0 e 1.
- Análise do ajuste do modelo:** testes de significância dos parâmetros e tabela de classificação (matriz de confusão).





Regressão Logística Multinomial

- Variável dependente se apresenta na forma **qualitativa com mais de duas categorias**.
- Por exemplo, para três possíveis respostas (*labels* 0, 1 ou 2, por exemplo), e sendo 0 a categoria de referência escolhida, teremos duas outras possibilidades de evento em relação a esta categoria (1 ou 2).
- Dessa forma, são definidos dois vetores de variáveis explicativas, com os respectivos parâmetros estimados (dois logitos):

$$Z_{1i} = \alpha_1 + \beta_{11}X_{1i} + \dots + \beta_{k1}X_{ki}$$

$$Z_{2i} = \alpha_2 + \beta_{12}X_{1i} + \dots + \beta_{k2}X_{ki}$$

- Logo, o **número de logitos estimados será ($M - 1$)**, sendo M o número de categorias de Y .

- Sendo $p_i = \frac{e^{Z_i}}{1 + e^{Z_i}}$ a **probabilidade de ocorrência do evento**, temos que:

- Probabilidade de ocorrência da **categoria 0** (referência): $P_{i_0} = \frac{1}{1 + e^{Z_{1i}} + e^{Z_{2i}}}$

- Probabilidade de ocorrência da **categoria 1**: $P_{i_1} = \frac{e^{Z_{1i}}}{1 + e^{Z_{1i}} + e^{Z_{2i}}}$

- Probabilidade de ocorrência da **categoria 2**: $P_{i_2} = \frac{e^{Z_{2i}}}{1 + e^{Z_{1i}} + e^{Z_{2i}}}$

Interpretação e Eficiência Global do Modelo Multinomial

- Como na regressão logística binária, deve-se avaliar o resultado do **teste χ^2** para o modelo de regressão logística multinomial, bem como os resultados dos **testes z** para os parâmetros estimados das variáveis preditoras.
- **Interpretação:** os parâmetros das variáveis devem ser analisados em relação à categoria de referência da variável dependente.
- **Eficiência do modelo:** a classificação das observações deve ser realizada a partir da maior probabilidade estimada para cada observação (aqui, ao contrário da regressão logística binária, não faz sentido a definição de um *cutoff*).



MUITO OBRIGADO!

Prof. Dr. Luiz Paulo Fávero

