Modelos de Regressão Logística

- TÉCNICA DE ANÁLISE MULTIVARIADA UTILIZADA PARA AFERIÇÃO DA PROBABILIDADE DE OCORRÊNCIA DE UM EVENTO E PARA IDENTIFICAÇÃO DAS CARACTERÍSTICAS DOS ELEMENTOS PERTENCENTES A CADA CATEGORIA ESTABELECIDA PELA DICOTOMIA DA VARIÁVEL DEPENDENTE (VARIÁVEL GRUPO).

- VARIÁVEL DEPENDENTE : BINÁRIA

- VARIÁVEIS INDEPENDENTES : MÉTRICAS OU NÃO MÉTRICAS

Exemplos de Aplicação

- MARKETING: DIFERENCIAR OS CONSUMIDORES LEAIS DOS NÃO LEAIS A UM PRODUTO / MARCA EM TERMOS DO PERFIL DEMOGRÁFICO.

- ADMINISTRAÇÃO BANCÁRIA: DIFERENCIAR OS CLIENTES ADIMPLENTES DOS INADIMPLENTES COM RELAÇÃO A EMPRÉSTIMOS BANCÁRIOS.

- EDUCAÇÃO: DIFERENCIAR ALUNOS COM CHANCE DE TERMINAR O CURSO DE PÓS-GRAD DAQUELES COM POUCAS POSSIBILIDADES.

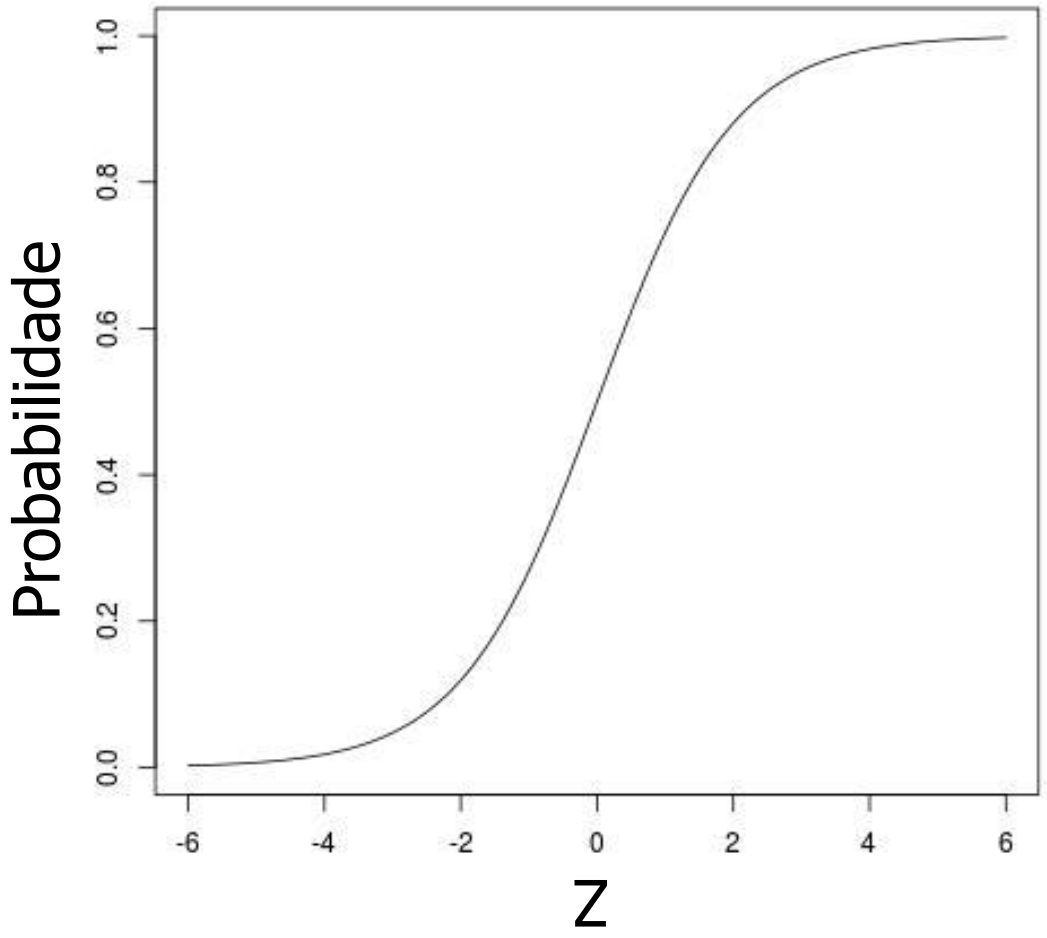
Objetivo

- ENCONTRAR UMA FUNÇÃO LOGÍSTICA, FORMADA POR MEIO DE PONDERAÇÕES DAS VARIÁVEIS (ATRIBUTOS), CUJA RESPOSTA PERMITA ESTABELECER A PROBABILIDADE DE OCORRÊNCIA DE DETERMINADO EVENTO E A IMPORTÂNCIA DAS VARIÁVEIS (PESO) PARA ESTA OCORRÊNCIA.

Diferenças Entre as Técnicas

- REGRESSÃO LINEAR
  - LINEARIDADE DO FENÔMENO MEDIDO
  - VARIAÇÃO CONSTANTE DOS ERROS
  - INDEPENDÊNCIA DOS ERROS
  - NORMALIDADE DA DISTRIBUIÇÃO DOS ERROS
  - PROBLEMA COM MULTICOLINEARIDADE
- ANÁLISE DISCRIMINANTE
  - LINEARIDADE DO FENÔMENO MEDIDO
  - DISTRIBUIÇÃO NORMAL DAS VARIÁVEIS INDEPENDENTES
  - GRUPOS COM MATRIZES DE COVARIÂNCIA IGUAIS
  - PROBLEMA COM MULTICOLINEARIDADE
- REGRESSÃO LOGÍSTICA
  - CURVA NA FORMA DE S NA RELAÇÃO DA VARIÁVEL DEPENDENTE COM CADA VARIÁVEL INDEPENDENTE
  - DISTRIBUIÇÃO NORMAL DAS VARIÁVEIS INDEPENDENTES (IMPACTO PEQUENO SE CONDIÇÃO NÃO SATISFEITA)

Representação Gráfica



Conceitos

- PROBABILIDADE
- DESIGUALDADE
- LOGITO
- COEFICIENTES LOGÍSTICOS

Conceitos

- **PROBABILIDADE**

SEJA Y A RESPOSTA A UM ESTÍMULO (SIM OU NÃO)

(pode ser preferência por um produto, adimplência, aprovação em um curso etc.)

p : probabilidade da resposta sim

1 - p : probabilidade da resposta não

Conceitos

- **DESIGUALDADE DE UMA RESPOSTA SIM**

$$\text{desigualdade} = \frac{p}{1-p}$$

$$p = \frac{\text{desigualdade}}{1+\text{desigualdade}}$$

- Exemplo : se p = 0,5 , desigualdade = 1
  se p = 0,75 , desigualdade = 3

Conceitos

- **LOGIT**

LOGARITMO NATURAL DE UMA DESIGUALDADE DE UMA RESPOSTA SIM

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

$$e^{\text{log it}} = e^{Z} = \frac{p}{1-p} = \text{desigualdade}$$

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

Modelo de Regressão Logística

- BASE DO MODELO : LOGITS = LOGARITMOS NATURAIS DAS DESIGUALDADES

- MÉTODO PARA PREVISÃO DOS LOGITS : MÉTODO DA MÁXIMA VEROSSIMILHANÇA

- $$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k + U$$

  $$\hat{Z} = B_0 + B_1 X_1 + B_2 X_2 + \ldots\ldots + B_k X_k$$

- B_1, B_2, ..., B_k: coeficientes logísticos

- PASSOS
  - obter a estimativa de Z
  - obter o valor da desigualdade
  - obter p

Exemplo

- Em um processo político de eleição, pretende-se discriminar as pessoas segundo a preferência por um determinado candidato X.

Y : preferência pelo candidato X
(0 = não, 1 = sim)

X_1 : idade em anos

X_2 : sexo (0 = masculino, 1 = feminino)

Exemplo

$$\hat{Z} = -10{,}83 + 0{,}28 \,.\, \mathbf{Idade} + 2{,}30 \,.\, \mathbf{Sexo}$$

- Características Sexo Masculino e 40 anos
  - Z estimado = 0,37
  - desigualdade = 1,448
  - p = 0,59
- Características Sexo Feminino e 40 anos
  - Z estimado = 2,67
  - desigualdade = 14,44
  - p = 0,94

Exemplo

- INTERPRETAÇÃO DOS COEFICIENTES
  - Idade : 0,28
  - Desigualdade : EXP (0,28) = 1,32

Exemplo

- EXEMPLO : Sexo Masculino e 41 anos
  - Z estimado = 0,65
  - desigualdade = 1,9155
  - desigualdade anterior = 1,448
  - fator de mudança na desigualdade=1,32

- Qual é o impacto no valor de p ?

$$p = \frac{\text{desiguald}^{1,32}}{1+\text{desiguald}^{1,32}} = \frac{1,448^{1,32}}{1+1,448^{1,32}}$$

$$p = 0,657$$

Procedimentos de Seleção das Variáveis Preditoras

- ENTER : todas as variáveis incluídas em um único passo.
- FORWARD STEPWISE : variáveis selecionadas em cada passo, segundo estatísticas de escores, com base em vários critérios: maior redução no valor de -2LL, maior coeficiente de Wald, maior probabilidade condicional de máxima verossimilhança.
- BACKWARD STEPWISE : variáveis eliminadas em cada passo, segundo estatísticas de escores, com base em vários critérios.

Testes de Hipóteses Sobre os Coeficientes

- ESTATÍSTICA DE WALD

Quadrado da razão entre o coeficiente e o seu erro padrão.

- ESTATÍSTICA COM DISTRIBUIÇÃO DE QUI-QUADRADO

- H_0 : o coeficiente é igual a zero.

- PROPRIEDADE INDESEJÁVEL DA ESTATÍSTICA WALD : quando o valor absoluto do coeficiente for grande, o erro padrão também o será, gerando baixo valor para esta estatística, sendo H_0 não rejeitada

Coeficiente de Correlação Parcial

$$R = \pm \sqrt{\frac{\text{estatística Wald} - 2k}{-2LL_{(0)}}}$$

k: g.l. para a variável

$-2LL_{(0)}$: medida de ajuste do modelo somente como intercepto

Se estatística Wald $< 2k \rightarrow R = 0$

Qualidade de Ajuste do Modelo

- GOODNESS-OF-FIT STATISTIC

Comparação das probabilidades observadas com as previstas pelo modelo.

$$Z^2 = \sum \frac{\text{Residual}_i^2}{P_i(1-P_i)}$$

Qualidade de Ajuste do Modelo

- MODEL CHI-SQUARE AND IMPROVEMENT

Teste equivalente ao teste F da regressão linear múltipla.

Improvement : diferença no valor de -2LL entre passos sucessivos da Regressão Logística Stepwise ou entre o modelo com todas as variáveis e somente com o intercepto.

O teste Qui-Quadrado testa a significância da mudança (improvement).

Qualidade de Ajuste do Modelo

- "PSEUDO R^2"

$$R^2_{logit} = \frac{-2LL_{null} - (-2LL_{model})}{-2LL_{null}}$$

Qualidade de Ajuste do Modelo

- HOSMER AND LEMESHOW GOODNESS-OF-FIT TEST

H_0: as classificações em grupo previstas são iguais às observadas

Qualidade de Ajuste do Modelo

- COX AND SNELL - R^2

Medida comparável ao R^2 da regressão linear múltipla

Qualidade de Ajuste do Modelo

- NAGELKERKE - R^2

Medida comparável ao R^2 da regressão linear múltipla.

Qualidade de Ajuste do Modelo

- TABELA DE CLASSIFICAÇÃO (MATRIZ DE CONFUSÃO)

Comparação dos valores observados da variável dependente com os previstos.

Se valor previsto < 0,5 , a observação é classificada a posteriori em Y = 0

Se valor previsto > 0,5 , a observação é classificada a posteriori em Y = 1

CURVA ROC

Sensibilidade

É a proporção de verdadeiros positivos: a capacidade do sistema em predizer corretamente a condição.

SENS = ACERTOS POSITIVOS / TOTAL DE POSITIVOS

= VP / (VP + FN)

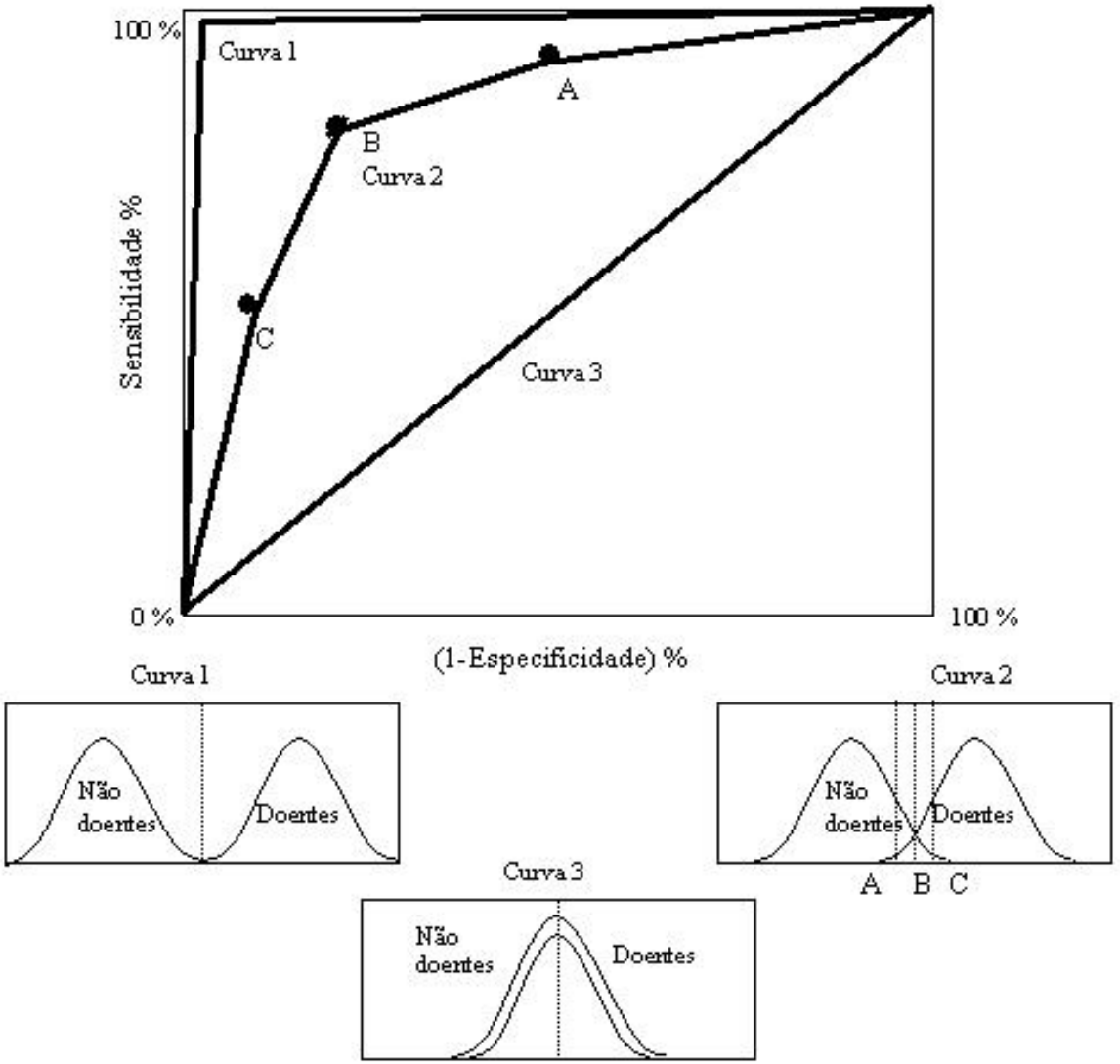
Especificidade

É a proporção de verdadeiros negativos: a capacidade do sistema em predizer corretamente a ausência da condição.

SPEC = ACERTOS NEGATIVOS / TOTAL DE NEGATIVOS

= VN / (VN + FP)

CURVA ROC


100 %
Curva 1
A
B
Curva 2
Sensibilidade %
C
Curva 3
0 %
100 %
(1-Especificidade) %
Curva 1
Não doentes
Doentes
Curva 2
Não doentes
Doentes
A B C
Curva 3
Não doentes
Doentes

Qualidade de Ajuste do Modelo

- VEROSSIMILHANÇA (L = Likelihood)

Probabilidade de obter os resultados da amostra, dadas as estimativas dos parâmetros do modelo logístico.

-2LL = medida da qualidade do ajuste

L	LL = log L	-2LL
1	0	0
0,7	-0,155	0,310
0,4	-0,398	0,796

Métodos de Diagnóstico

- RESIDUAL : DIFERENÇA ENTRE A PROBABILIDADE DO EVENTO OBSERVADA E A PROBABILIDADE PREVISTA

- STANDARDIZED RESIDUAL

$$Z_i = \frac{\text{Residual}_i}{\sqrt{P_i(1-P_i)}}$$

- DEVIANCE

$$\text{Deviance} = -\sqrt{-2\log(\text{probab prev para gpo observ})}$$

- STUDENTIZED RESIDUAL : PARA CADA CASO É A MUDANÇA EM DEVIANCE SE O CASO É EXCLUIDO

EXEMPLO 2

- O Departamento de Marketing de uma empresa de cartões de crédito pretende lançar uma campanha para que seus usuários com padrão standard mudem para um padrão mais elevado, oferecendo um desconto para a taxa anual do novo cartão.

Para uma amostra de 100 clientes com o padrão standard foram obtidas as variáveis:

Y : mudaria para o novo cartão (0 = não, 1 = sim)

X_1 : total de gastos no ano anterior

X_2 : possui cartão adicional (0 = não, 1 = sim)

- Deseja-se uma estimativa de compra do novo cartão para um cliente com gastos de US$ 38 mil e com um cartão adicional.

EXEMPLO 2

Beginning Block Number 0. Initial Log Likelihood Function					
-2 Log Likelihood 131,79114					
* Constant is included in the model.					

EXEMPLO 2

Method: Forward Stepwise (COND)							
Variables in the Equation							
Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
CART(1)	2,4415	1,1409	4,5794	1	,0324	,1399	11,4903
GAST	,5087	,1133	20,1589	1	,0000	,3712	1,6632
Constant	-14,1824	3,5055	16,3677	1	,0001		

EXEMPLO 2

-2 Log Likelihood (-2LL)	53,576
Goodness of Fit	65,666
Cox & Snell - R^2	,543
Nagelkerke - R^2	,741

EXEMPLO 2

	Improv.			Correct				
Step	Chi-Sq.	df	sig	Chi-Sq.	df	sig	Class %	Variable
1	72,321	1	,000	72,321	1	,000	87,00	IN: GAST
2	5,894	1	,015	78,215	2	,000	90,00	IN: CART

EXEMPLO 2

Hosmer and Lemeshow Goodness-of-Fit Test					
Y	= não	Y	= sim		
Group	Observed	Expected	Observed	Expected	Total
1	11,000	10,915	,000	,085	11,000
2	11,000	10,120	,000	,880	11,000
3	7,000	7,774	3,000	2,226	10,000
4	4,000	4,283	6,000	5,717	10,000
5	2,000	2,158	8,000	7,842	10,000
6	1,000	1,033	9,000	8,967	10,000
7	1,000	,506	9,000	9,494	10,000
8	,000	,179	11,000	10,821	11,000
9	,000	,028	10,000	9,972	10,000
10	,000	,003	7,000	6,997	7,000
		Chi-Square	df	Significance	
Goodness-of-fit test		2,1578	8	,9758	

EXEMPLO 2

	Grupo previsto		
Grupo observado	não	sim	% previsões corretas
não	31	6	83,78%
sim	4	59	93,65%
% previsões corretas	88,57%	90,77%	**90,00%**

EXEMPLO 2

```
 Observed Groups and Predicted Probabilities

        32 +                                                            +
           I                                                            I
           I                                                            I
F          I                                                           sI
R       24 +                                                           s+
E          I                                                           sI
Q          I                                                           sI
U          I                                                           sI
E       16 +                                                           s+
N          I                                                           sI
C          In                                                          sI
Y          In                                                          sI
         8 +n                                                          s+
           In      n                                               s   ssI
           In    n n    s                 s           s          ss s sssI
           Innn n nn   nn    s    n n       n   sn   s s n   ss nsss nssnsnsssI
Predicted ---------------+---------------+---------------+---------------
  Prob:   0             ,25             ,5              ,75             1
  Group:  nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnssssssssssssssssssssssssssssssss

          Predicted Probability is of Membership for sim
          The Cut Value is ,50
          Symbols: n - não
                   s - sim
          Each Symbol Represents 2 Cases.
```