Regressão Logística

ACH2036 – Métodos Quantitativos para Análise Multivariada
Prof. Regis Rossi A. Faria

EACH
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

USP

2º sem. 2022

Programa

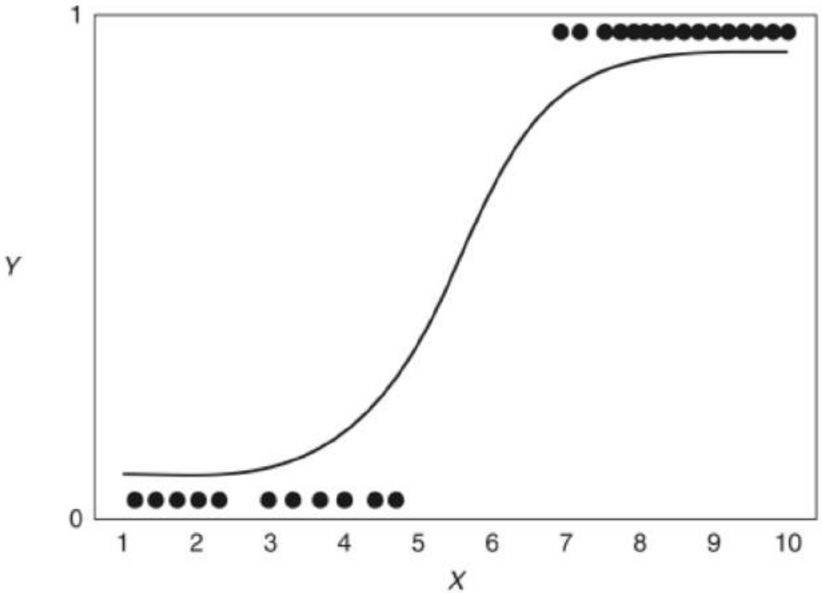
- Introdução (histórico, aplicabilidade)
- Modelo (equações, propriedades)
- Características (vantagens, suposições requeridas)
- Avaliação do modelo
- Exemplos

Introdução

- Regressão logística é um método usado para prever a probabilidade de ocorrência de valores de variáveis dependentes binárias (categóricas ou não-métricas) a partir de variáveis independentes (métricas e não-métricas)

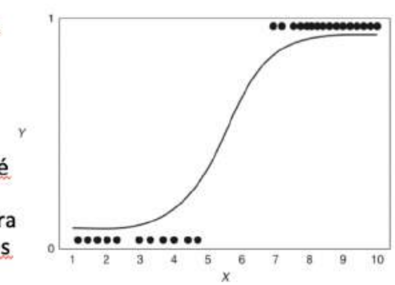
Introdução

- Regressão logística é um método usado para prever a probabilidade de ocorrência de valores de variáveis dependentes binárias (categóricas ou não-métricas) a partir de variáveis independentes métricas e/ou não-métricas
- A variável dependente *Y* assume 2 valores somente, mas o que fazemos é representar graficamente a *probabilidade de ocorrência P(Y)* contra os valores das variáveis independentes *X* por meio de uma curva em S (não-linear), em que *P(Y)* está restrita a um domínio entre 0 e 1



Introdução

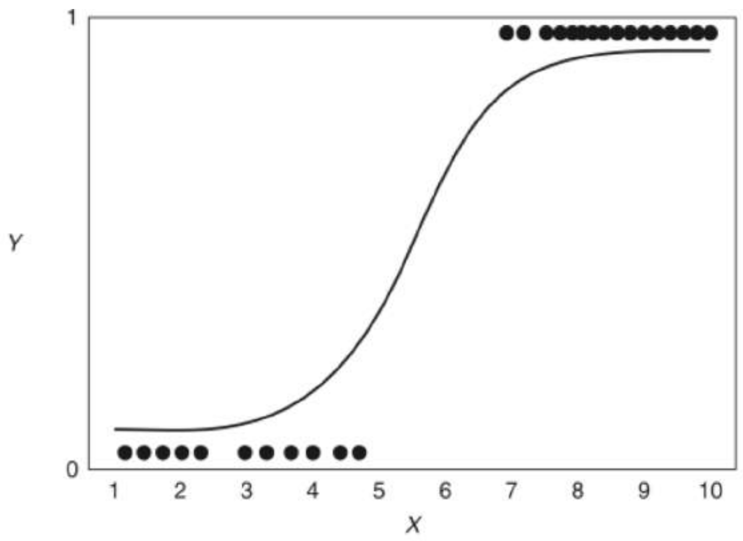
- Regressão logística é um método usado para prever a probabilidade de ocorrência de valores de variáveis dependentes binárias (categóricas ou não-métricas) a partir de variáveis independentes (métricas e não-métricas)
- A variável dependente *Y* assume 2 valores somente, mas o que fazemos é representar graficamente a *probabilidade de ocorrência P(Y)* contra os valores das variáveis independentes *X* por meio de uma curva em S (não-linear), em que *P(Y)* está restrita a um domínio entre 0 e 1



- Esta é uma situação muito comum:
  - ✓ ter uma variável que se quer prever de natureza binária (é ou não é) mas
  - ✓ que seu estado (de ser ou não ser) precise ser interpretado por meio de probabilidades (de ser ou não ser)
- Exemplos: Interesse em...
  - ✓ Saber o risco (probabilidade) de ter um acidente (variável binária)
  - ✓ Saber se um contrato pode (probabilidade) ser cancelado (situação binária)
  - ✓ Saber se um paciente pode (probabilidade) enfartar (ou enfarta ou não enfarta)

Introdução

- O modelo que relaciona as variáveis independentes x1, x2, ... com a variável dependente y (que se quer prever) parte de um modelo de regressão linear mas que se relaciona com uma quantidade nomeada ***logit*** = logaritmo (natural) de uma *razão de chances* (também chamada de razão de desigualdades)

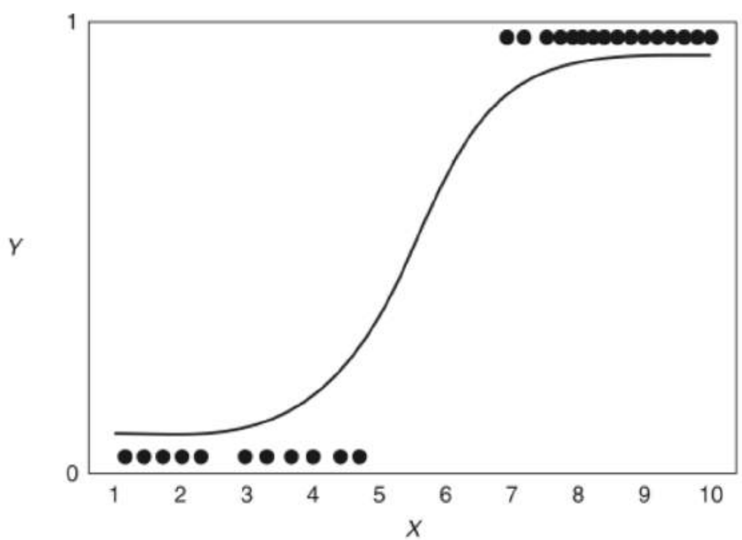


razão de chances

$$\ln\left(\frac{Y}{1-Y}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots$$

Introdução

- O modelo que relaciona as variáveis independentes x1, x2, ... com a variável dependente y (que se quer prever) parte de um modelo de regressão linear mas que se relaciona com uma quantidade nomeada *logit* = logaritmo (natural) de uma *razão de chances* (também chamada de razão de desigualdades)



parte linear do modelo

$$\ln\left(\frac{Y}{1-Y}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots$$

medida independente

logit, onde Y é a probabilidade de ocorrer o evento binário

medida dependente

Transformação da variável dependente

- A regressão logística deriva seu nome do uso da transformação *logit* usada sobre a variável dependente Y
- A equação

$$\mathrm{logit(Y)} = \ln\left(\frac{Y}{1-Y}\right) = b_0 + b_1X_1 + b_2X_2 + \cdots$$

tendo que *Y = P(evento)* pode ser reescrita equivalentemente como

$$\frac{Prob(evento)}{1-Prob(evento)} = e^{b_0+b_1X_1+b_2X_2+\cdots}$$

razão de chances (*odds ratio*)

Transformação da variável dependente

razão de chances (*odds ratio*)

- Com

$$\frac{P(evento)}{1-P(evento)} = \frac{Y}{1-Y} = e^{b_0+b_1X_1+b_2X_2+\cdots}$$

- A probabilidade do evento $P(evento) = Y$ pode ser expressa por

$$P(evento) = Y = \frac{e^{(b_0+b_1X_1+b_2X_2+\cdots)}}{1+e^{(b_0+b_1X_1+b_2X_2+\cdots)}}$$

ou

$$P(evento) = Y = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\cdots)}}$$

Efeitos dos coeficientes sobre *P*(*evento*)

razão de chances (*odds ratio*)

- Como

$$\frac{P(evento)}{1-P(evento)} = \frac{Y}{1-Y} = e^{b_0+b_1X_1+b_2X_2+\cdots} = e^{b_0}.\, e^{b_1X_1} \ldots$$

- Notamos que o efeito do coeficiente b_1 sobre o a razão de chances é proporcional a e^{b_1}
  - Ex: se $b_1 = 0{,}3$ então a razão de chances (*rc*) será aumentada em $e^{b_1} \cong 1{,}35$ cada vez que X_1 aumentar de 1 unidade, isto é, um aumento de 35%
- Com $\frac{Y}{1-Y} = rc$ → $y = \frac{rc}{1+rc}$. Uma nova probabilidade Y' devido às chances estarem α vezes maiores, poderá ser expressa por $y' = \frac{\alpha.rc}{1+\alpha.rc} = \frac{\alpha.Y}{1+(\alpha-1)Y}$
  - Ex: para α=1,35 , se Y=0,8 então $Y' = \frac{1{,}35.0{,}8}{1+(0{,}35)0{,}8} = 0{,}8438$, isto é, *P(evento)* passou para 84,38% → um aumento de 4,38% e não de 35%

Efeitos dos coeficientes sobre *P*(*evento*)

razão de chances (*odds ratio*)

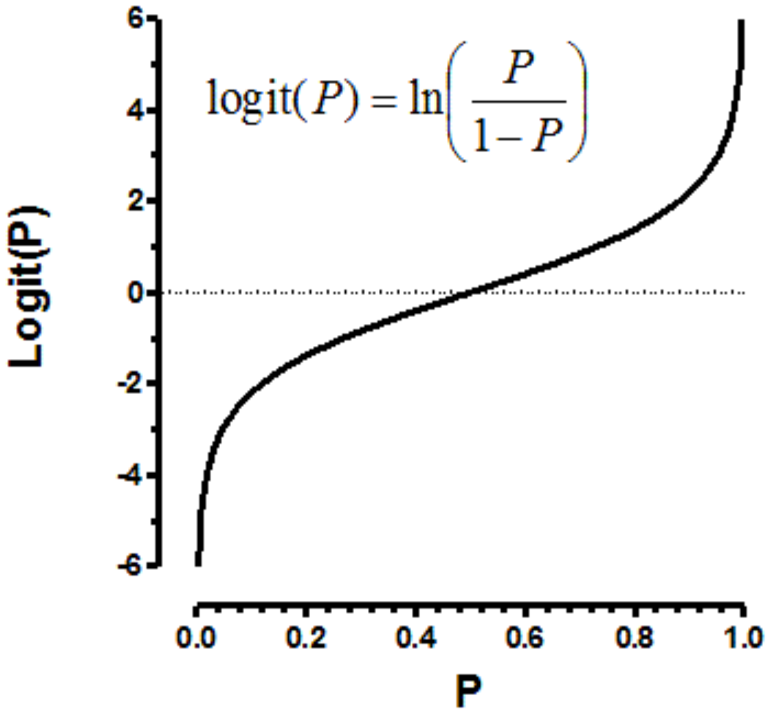
- Como

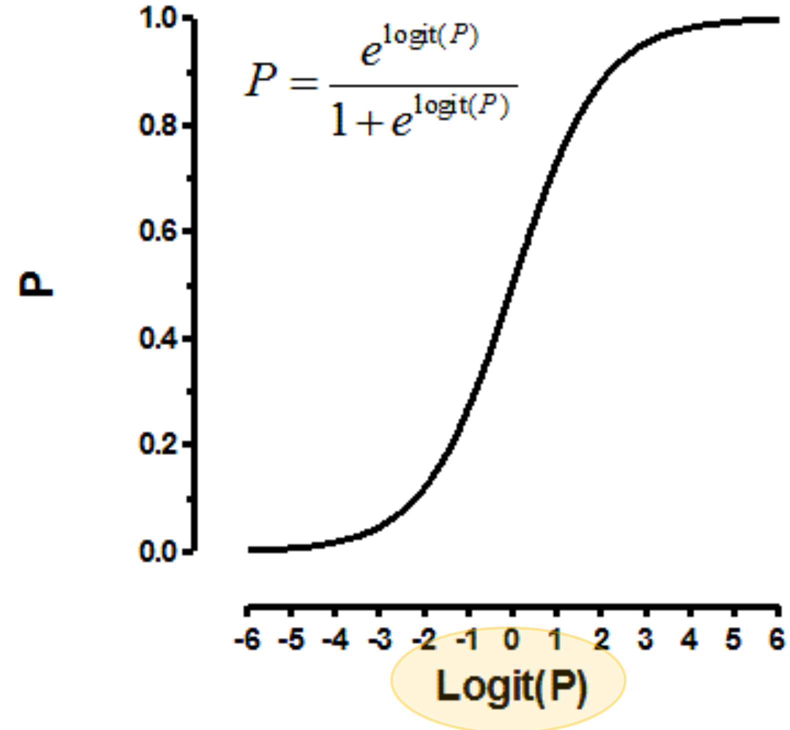
$$\frac{P(evento)}{1-P(evento)} = \frac{Y}{1-Y} = e^{b_0+b_1X_1+b_2X_2+\cdots} = e^{b_0}.\, e^{b_1X_1}\ldots$$

- Notamos que o efeito do coeficiente b_1 sobre o a razão de chances é proporcional a e^{b_1}
  - Ex: se $b_1 = 0{,}3$ então a razão de chances (rc) será aumentada em $e^{b_1} \cong 1{,}35$ cada vez que X_1 aumentar de 1 unidade, isto é, um aumento de 35%
- Com $\frac{Y}{1-Y} = rc \rightarrow y = \frac{rc}{1+rc}$. Uma nova probabilidade Y' devido às chances estarem α vezes maiores, poderá ser expressa por $y' = \frac{\alpha.rc}{1+\alpha.rc} = \frac{\alpha.Y}{1+(\alpha-1)Y}$
  - Ex: para α=1,35 , se Y=0,8 então $Y' = \frac{1{,}35.0{,}8}{1+(0{,}35)0{,}8} = 0{,}8438$, isto é, *P(evento)* passou para 84,38% → um aumento de 4,38% e não de 35%

Isto mostra que as variações de probabilidades não são lineares

P (=Y) e logit(P)





$$b_0 + b_1X_1 + b_2X_2 + \cdots$$

Fonte imagem: http://estatsite.com.br/2018/08/29/regressao-logistica-conceitos-e-formula/

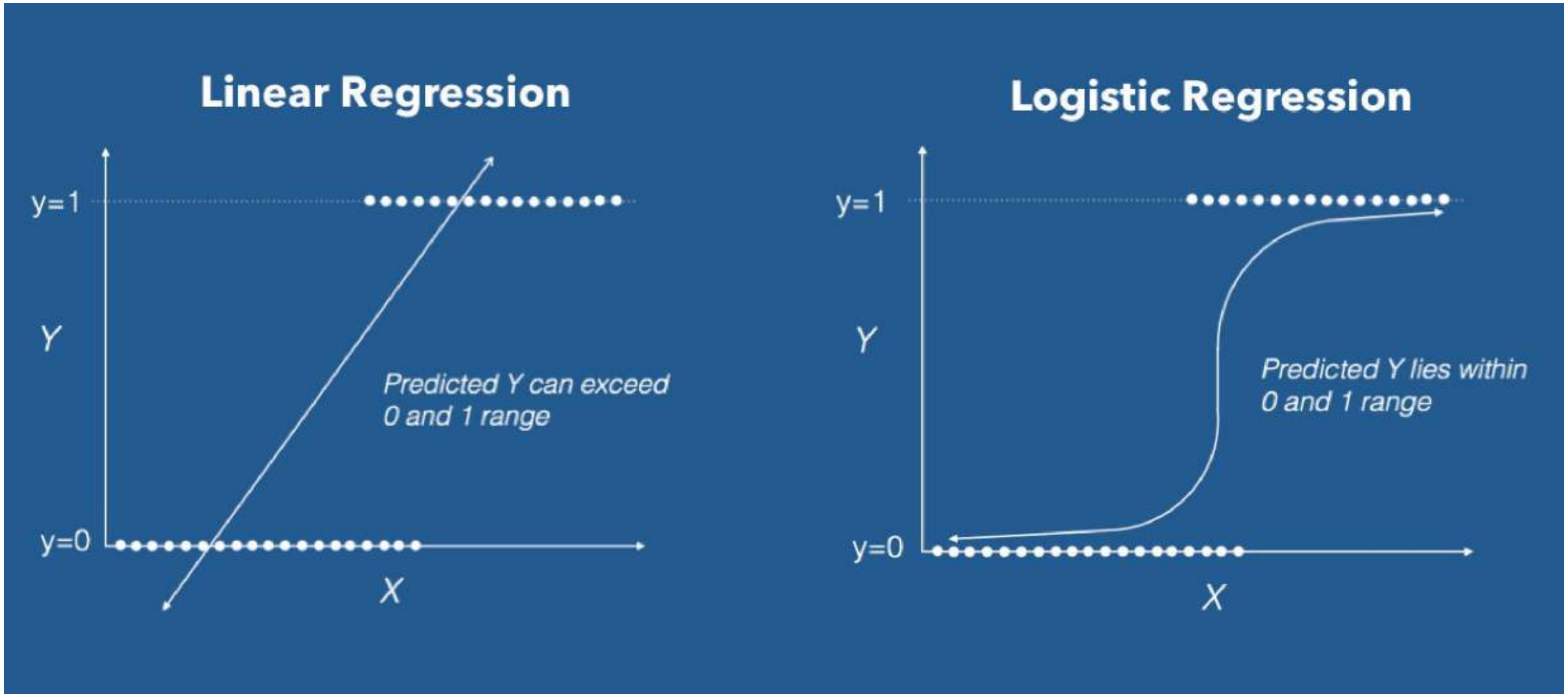
Valores comparados

- Probabilidades variam de 0 a 1
- Razão de chances varia de 0 a +∞ (NC) (crescente)
- *logit* varia numa faixa entre –∞ (NC) e +∞ (NC), passando por 0 quando p=0,5 e a razão de chances = 1,0

Probabilidade	Razão de desigualdades	Logaritmo (Logit)
0,00	0,00	NC
0,10	0,111	–2,197
0,30	0,428	–0,847
0,50	1,000	0,000
0,70	2,333	0,847
0,90	9,000	2,197
1,00	NC	NC

NC = Não pode ser calculado

Faixas dos valores: *linear* X *logística*



Onde Y é a variável que nós queremos prever

Fonte imagem: https://www.machinelearningplus.com

Parte operacional

- O trabalho com a regressão logística, semelhante com outras técnicas, envolve
  - ✓ estimar os coeficientes logísticos b
  - ✓ estimar a variável estatística Y
  - ✓ avaliar a adequação do modelo (o ajuste do modelo)
  - ✓ interpretar os resultados (coeficientes)

- Estimando a pertinência a um grupo:
  - Para cada observação com valores ***X***, a técnica prevê uma probabilidade 0<***Y***<1, usando os coeficientes ***b*** estimados
  - ✓ Se Y>0,5 → Y=1
  - ✓ Se Y≤0,5 → Y=0

Regressão Logística

Em geral, modelos de regressão não linear são usados em duas situações: casos em que as variáveis respostas são qualitativas e os erros não são normalmente distribuídos.

O modelo de regressão não linear logístico binário é utilizado quando a variável resposta é qualitativa com dois resultados possíveis. Por exemplo:

- *Sobrepeso de crianças* → tem sobrepeso ou não tem sobrepeso
- Solvência de empresa → tem ativo maior que passivo ou não
- Esta variável terá assumida uma distribuição binomial.

Este modelo pode ser estendido quando a variável resposta qualitativa tem mais do que duas categorias; por exemplo, a pressão sanguínea pode ser classificada como alta, normal e baixa.

Modelos de regressão com variáveis respostas binárias

Em muitos estudos a variável resposta tem duas possibilidades e, assim, pode ser representada pela variável indicadora, recebendo os valores 0 (zero) e 1 (um).

Exemplos:

1) O objetivo da análise é verificar a proporção de óbitos neonatais com função da mãe ter diabetes *mellitus* tipo 1. A variável resposta tem duas possibilidades: a criança morreu ou não. Estes resultados podem ser codificados como 1 e 0 (de acordo com o interesse).

Modelos de regressão com variáveis respostas binárias

Em muitos estudos a variável resposta tem duas possibilidades e, assim, pode ser representada pela variável indicadora, recebendo os valores 0 (zero) e 1 (um).

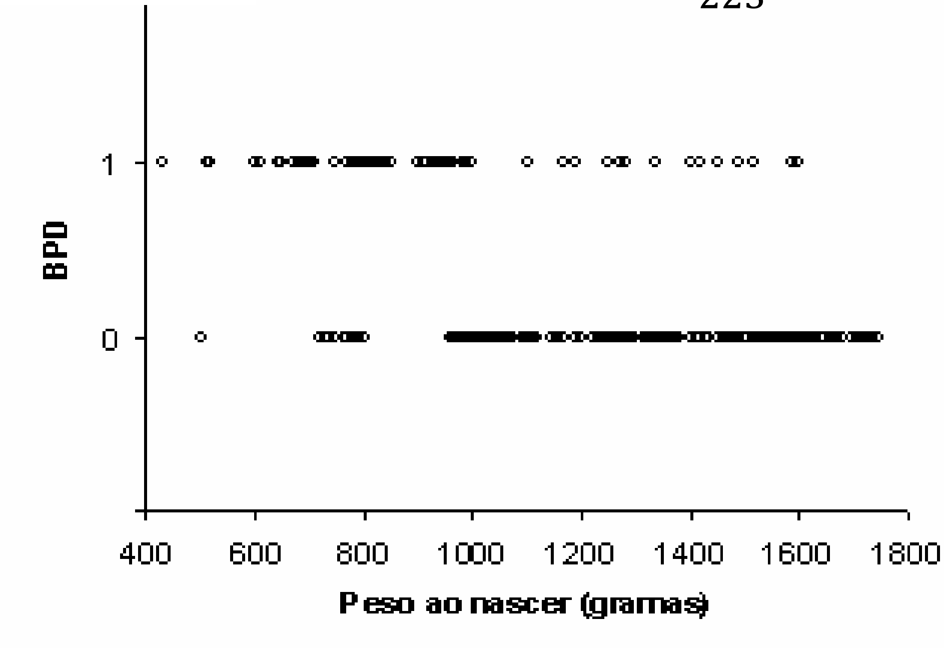
Exemplos:

2) Num estudo sobre a participação das esposas no mercado de trabalho, como função da idade da esposa, número de filhos e rendimento do marido, a variável resposta Y foi definida do seguinte modo: *a mulher participa no mercado de trabalho ou não*. Novamente, estas respostas podem ser codificadas como 1 e 0, respectivamente.

Exemplo:
Peso de recém-nascidos *versus* BPD

- Bebês ao nascer abaixo de 1750 gramas ficam confinados em uma UTI neonatal. Em uma amostra de 223 bebês, 76 apresentaram diagnóstico de BPD (displasia broncopulmonar). A probabilidade de uma criança, nestas condições, ter BPD é

$$p = \frac{76}{223} = 0{,}341$$



Exemplo do *dataset*

Bebê	BPD	peso
1	1	500
2	1	600
3	0	1000
...	...	...
223	...	...

Modelo Geral

Conceitos

Modelo de Regressão Múltipla

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$$

Objetivo: Estimar o valor médio da resposta, considerando algumas variáveis explicativas

$$E(Y_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Interpretação da função de resposta quando a variável resposta é binária

Vamos considerar o modelo de regressão linear simples:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}$$

A resposta esperada é dada por:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

Função Logística

- Do modelo geral $E(Y_i) = \beta_0 + \beta_1 X_i$
- Partimos para um modelo inicial probabilístico $p = \beta_0 + \beta_1 x$
- Mas neste modelo a variável dependente (p) pode assumir valor <0 e >1. Para contornar esta limitação, efetua-se uma *transformação logística* na variável dependente, tal que a função (ou equação) logística fica

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad \text{em que } 0 < p < 1$$

Função Logística

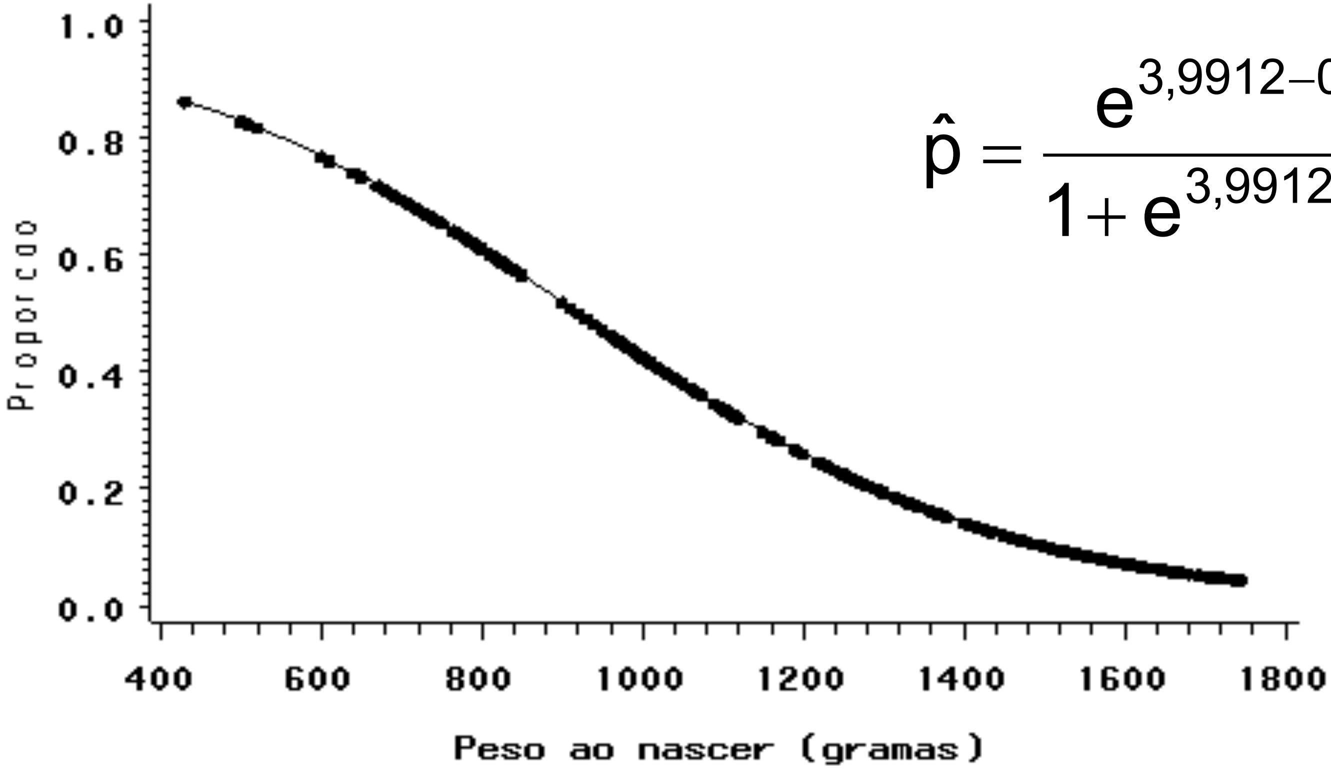
- Do modelo geral $E(Y_i) = \beta_0 + \beta_1 X_i$
- Partimos para um modelo inicial probabilístico $p = \beta_0 + \beta_1 x$
- Mas neste modelo a variável dependente (p) pode assumir valor <0 e >1. Para contornar esta limitação, efetua-se uma *transformação logística* na variável dependente, tal que a função (ou equação) logística fica

$$p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

em que $0 < p < 1$

nova variável representa a probabilidade de ocorrência do evento, isto é, *p(evento)*

Aplicando ao exemplo dos bebês

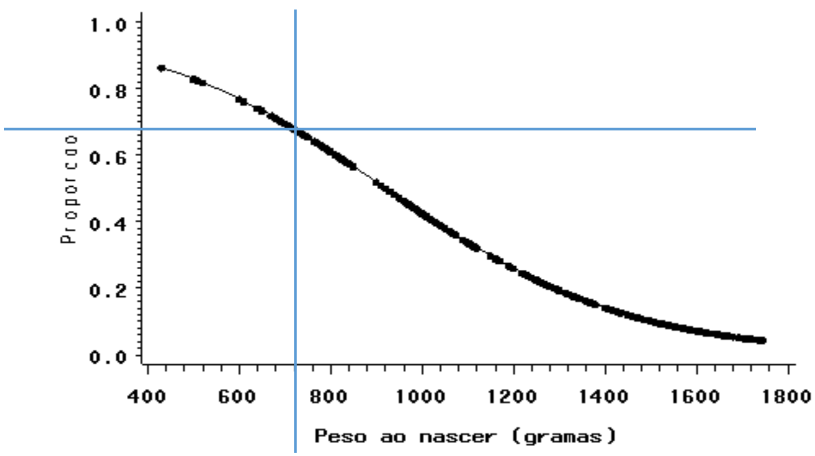


$$\hat{p} = \frac{e^{3,9912 - 0,0043\,x}}{1 + e^{3,9912 - 0,0043\,x}}$$

Aplicando ao exemplo dos bebês

Para encontrar a probabilidade de que uma criança que pesa 750 gramas no nascimento desenvolva BPD, substitui-se o valor x=750 na função.

$$\hat{p} = \frac{e^{3,9912 - 0,0043\,(750)}}{1 + e^{3,9912 - 0,0043\,(750)}} = 0,6827$$



Dados Categorizados

- Consideremos que categorizamos os casos de BPD por 3 faixas de peso, como a seguir

Fator: o peso de nascimento do bebê (0 |-- 950, 950 |-- 1350, 1350 |-- 1750)
Variável resposta: o bebê *está* ou *não está* doente com BPD

Peso ao nascer (gramas)	Tamanho da amostra	Quantidade com BPD	p
0 |-- 950	68	49	0,721
950 |-- 1350	80	18	0,225
1350 |-- 1750	75	9	0,120
	223	76	0,341

Dados Categorizados

- Consideremos que categorizamos os casos de BPD por 3 faixas de peso, como a seguir

Fator: o peso de nascimento do bebê (0 |-- 950, 950 |-- 1350, 1350 |-- 1750)
Variável resposta: o bebê está ou não está doente com BPD

	Peso ao nascer (gramas)	Tamanho da amostra	Quantidade com BPD	p
$X_1 \rightarrow$	0 |-- 950	68	49	0,721
$X_2 \rightarrow$	950 |-- 1350	80	18	0,225
	1350 |-- 1750	75	9	0,120
		223	76	0,341

Vamos introduzir 2 variáveis categóricas "dummies" para refletir o pertencimento a estas faixas

Dados Categorizados

- O *dataset* com X_1 e X_2 fica:

Bebê	BPD	peso	X1	X2
1	1	500	1	0
2	1	600	1	0
3	0	1000	0	1
4		1450	0	0
...	...	...	...	...
223	...	...	...	...

Modelo de Regressão Logística para os dados

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}$$

$$p = \frac{e^{-1,992 + 2,940 X_1 + 0,756 X_2}}{1 + e^{-1,992 + 2,940 X_1 + 0,756 X_2}}$$

X_1 representa o peso de 0 a 950 gramas e X_2 o peso de 950 a 1350 gramas

Modelo de Regressão Logística para os dados

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}$$

$$p = \frac{e^{-1,992 + 2,940 X_1 + 0,756 X_2}}{1 + e^{-1,992 + 2,940 X_1 + 0,756 X_2}}$$

X_1 representa o peso de 0 a 950 gramas e X_2 o peso de 950 a 1350 gramas

Se:

e^{bi} =1, então a chance de apresentar y=1 é a mesma da classe [1350-1750)

e^{bi} >1, então a chance de apresentar y=1 é maior que da classe [1350-1750)

e^{bi} <1, então a chance de apresentar y=1 é menor que da classe [1350-1750)

Regressão logística

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}$$

Interpretamos e^{β_1}, ..., e^{β_k}
como fatores de razão de chances
(odds ratio)

Regressão logística

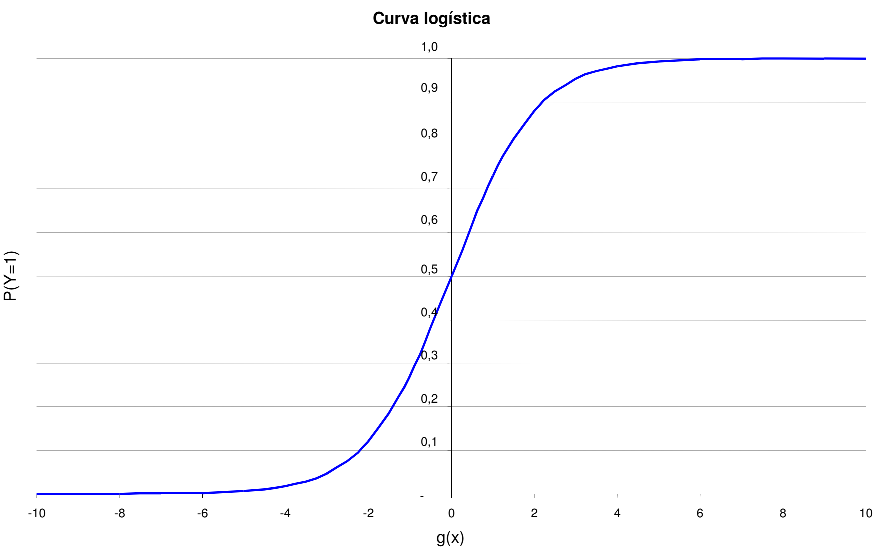
- A função

$$p = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}} =$$

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2)}}$$

é da forma geral

$$p = \frac{1}{1+e^{-g(x)}}$$



- Quando $g(x) \rightarrow +\infty$ então $p(Y=1) \rightarrow 1$, e
- Quando $g(x) \rightarrow -\infty$ então $p(Y=1) \rightarrow 0$

Analisando a relação entre as variáveis no exemplo

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	PESO			53.748	2	.000	
	PESO(1)	2.940	.446	43.364	1	.000	18.912
	PESO(2)	.756	.445	2.885	1	.089	2.129
	Constant	-1.992	.355	31.441	1	.000	.136

a. Variable(s) entered on step 1: PESO.

Qual a interpretação da significância de PESO (1) e PESO (2) ?

Qual a interpretação do exp(b) ?

Analisando a razão de chances (rc)

Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
18,912	7,884	45,368
2,129	,890	5,092
,136		

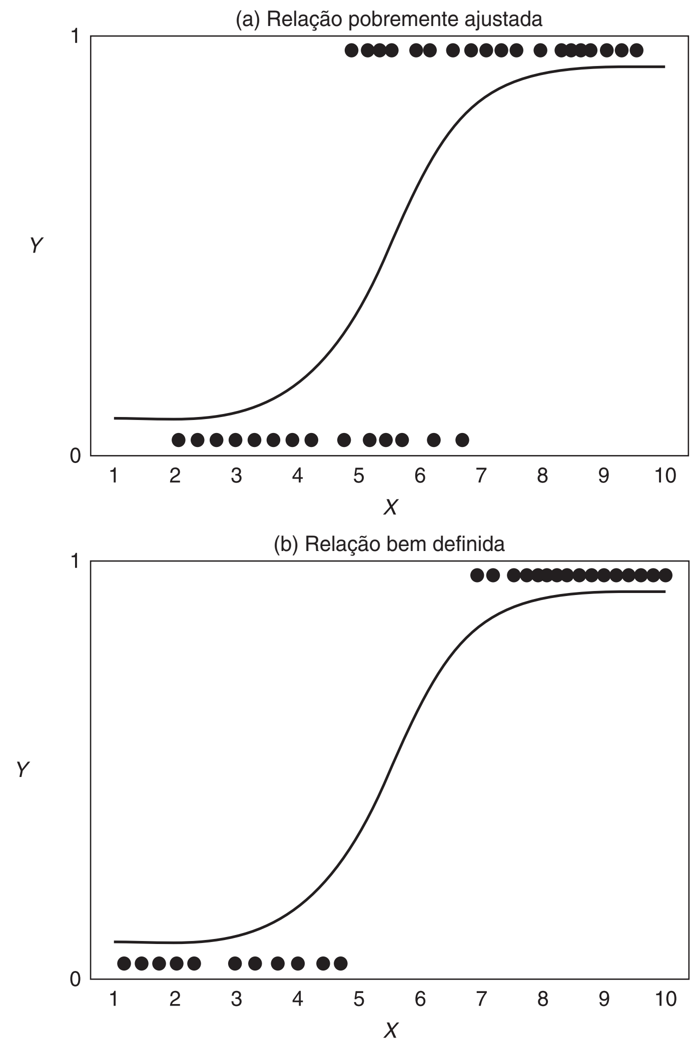
Interpretação: A chance de uma criança com peso entre 0 e 950 gramas ter a presença da BPD é 18,9 vezes maior do que uma criança com peso entre 1350 e 1750 gramas

Suposições do modelo logístico

- O modelo logístico admite variáveis independentes métricas e não métricas , apresenta pequeno número de suposições, é mais flexível, robusto, contorna bem restrições de outros modelos multivariados, e isto é um dos motivos por que é muito utilizado
- Não requer
  - Homogeneidade de variância ou variância constante (homoscedasticidade)
  - Normalidade na distribuição dos erros
  - Normalidades das variáveis independentes
- Requer
  - Valor esperado do erro = 0
  - Inexistência de autocorrelação entre erros; e entre estes e as variáveis independentes
  - Ausência de multicolinearidade perfeita entre as variáveis independentes

Medidas de avaliação do modelo

- Ajuste do modelo
  - É importante verificar se o modelo extraído atinge a performance suficiente e adequada para previsão
  - Checar o ajuste da curva logística aos dados
  - Verificar a significância dos coeficientes → verificar se podem ser usados como estimadores de probabilidade (medição do quanto os coeficientes das variáveis independentes explicam das variações na probabilidade)
  - Verificar a taxa de erro, acurácia do preditor



Medidas de avaliação do modelo

- Na regressão linear usamos a estatística F e o coeficiente de determinação R^2 para testar a significância e poder explicativo do modelo, mas em regressão logística o método de estimação dos coeficientes é o da máxima verossimilhança (e não dos mínimos quadrados, que produz R^2) portanto precisamos de outras medidas para avaliar o modelo
- Medidas usadas:
  - Log Likelihood value
  - R^2 do modelo logístico
- Testes usados:
  - Hosmer e Lemeshow
  - Teste Wald
  - Teste Cox-Snell (pseudo R^2)

Medidas de avaliação do modelo

- *Log Likelihood value* (valor de verossimilhança)
  - Papel parecido com o da estatística F
  - Notação: -2LL (= logaritmo natural do likelihood value *-2)
  - Nível ideal: 2LL=0 (ajuste perfeito)
    - Quando LL=1 → -2*ln(LL) = -2*ln(1) = 0
    - Se a probabilidade máxima de um evento ocorrer é =1, quanto mais próximo de zero o LL maior será o poder preditivo do modelo
  - -2LL não é passível de interpretação isoladamente, mas sim confrontado com uma base de referência (ex: ao comparar desempenho de modelos alternativos)
  - Serve para verificar se o modelo melhora com a inclusão/exclusão de alguma variável independente

Medidas de avaliação do modelo

- R^2 do modelo logístico
  - Trata-se de um pseudo-R^2
  - R^2logit pode calculado da seguinte forma: $$R^2_{\text{LOGIT}} = \frac{-2LL_{\text{nulo}} - (-2LL_{\text{modelo}})}{-2LL_{\text{nulo}}}$$
  - que expressa a variação percentual entre
    o LLvalue nulo (considerando apenas a constante)
    e o LLvalue do modelo (incorporando as variáveis explicativas)
  - Para -2LLmodelo = 0, teremos que o ajuste do modelo será perfeito

Medidas de avaliação do modelo

- Testes usados:
  - *Hosmer e Lemeshow*: é um teste qui-quadrado que consiste em dividir o número de observações em 10 classes, e então comparar as frequências preditas com as observadas →checa se há diferenças significativas entre as classificações do modelo e a realidade (observada)
  - A certo nível de significância (ex: 5%) busca-se aceitar a H0 (hipótese nula) de que não haja diferenças entre os valores preditos e observados → portanto quanto maior o nível de significância, melhor (ex: 0,1 é melhor que 0.05)

Medidas de avaliação do modelo

- Testes usados:
  - *Teste Wald*: afere o grau de significância de cada coeficiente da equação logística (inclusive a constante) → checa se cada parâmetro estimado é significativamente diferente de 0 (papel semelhante ao de um teste t, ao testar a hipótese de que um determinado coeficiente é nulo)
    - Estatística Wald: segue uma distribuição qui-quadrada
    - Wald = $(b/SE)^2$, onde b é o valor do coeficiente e SE = erro padrão

Medidas de avaliação do modelo

- Testes usados:
  - Cox-Snell R^2 : teste comparável ao R^2 da regressão linear → seu resultado expressa o percentual com que as variações ocorridas no Log da razão de chances são explicadas pelo conjunto das variáveis independentes → pode ser usado para comparar modelos concorrentes (prefira os valores que tem Cox-Snell mais elevado)
  - *Nagelkerke R2* : finalidade similar ao de Cox-Snell, só que normalizada entre 0 e 1

Resumo das características do método

- Os valores de Y estão restritos entre 0 e 1 (não saem deste domínio, como qualquer valor de probabilidade)
- Equivalente a uma análise discriminante com dois grupos
- A variável resposta tem distribuição de probabilidade binomial
- Admite, simultaneamente, variáveis independentes métricas e não-métricas
- Menos restritiva quanto a suposições iniciais impostas aos dados
- Atraente para aplicações de *machine learning*
- Facilidade em predizer a ocorrência de fenômenos em diversas áreas do conhecimento (ex: administração, sociologia, medicina) identificando a que grupo certos objetos, pessoas ou fenômenos pertencem

Resumo das características do método

- $logit(p) = \ln(p/(1-p))$
- $\ln(p/(1-p)) = \ln(razão\_chances)$
- $razão\_chances = p/ (1-p)$
- $p = 1/(1+e^{-g(x)})$
- Na regressão logística não assumimos uma relação linear entre a variável dependente e independente
- Erros não têm distribuição normal
- Utilizamos a máxima verossimilhança para estimar os coeficientes, e não mínimos quadrados

Exemplo: Interpretando o impacto de uma variável

- *logit* = $0{,}25x_1 + 0{,}4x_2$
  - x_1 = renda familiar
  - x_2 = no. de filhos
- p = probabilidade de alugar um imóvel
- Inicialmente: p = 0,3. Mas o casal ganhou um filho
  - a chance de alugar um imóvel era p/(1-p)=0,3/0,7=0,43
  - com mais um filho, a chance varia de $e^{0,4} = 1{,}49$
  - a razão de chance aumenta → 1,49*0,43=0,64
  - logo p passou para p'=0,39 (p~0,4) → um aumento de ~10%

Exemplo no R

Estudo sobre estado de adimplência ou não de clientes (status ST) que tenham renda mensal R, ND dependentes e estejam empregados (VE)

- Regressão logística no RStudio
- Exemplo de coeficientes obtidos:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.4776     1.6569   0.892 0.372501
R            -1.8824     0.4885  -3.853 0.000117 ***
ND            0.8596     0.3857   2.228 0.025854 *
VE            2.8221     0.8521   3.312 0.000926 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Realizando previsões com o modelo:

$$P(evento) = \frac{1}{1+e^{-(1{,}478-1{,}882R+0{,}860ND+2{,}822VE)}}$$

Exemplo no R

Estudo sobre estado de adimplência ou não de clientes (status ST) que tenham renda mensal R, ND dependentes e estejam empregados (VE)

- Regressão logística no RStudio
- Exemplo de coeficientes obtidos:

Erro padrão

Significância dos coeficientes

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.4776     1.6569   0.892 0.372501
R            -1.8824     0.4885  -3.853 0.000117 ***
ND            0.8596     0.3857   2.228 0.025854 *
VE            2.8221     0.8521   3.312 0.000926 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 126.450  on 91  degrees of freedom
Residual deviance:  50.307  on 88  degrees of freedom
AIC: 58.307

Number of Fisher Scoring iterations: 6
```

Desvio quando só a constante está no modelo (intercepto)

Soma dos quadrados dos resíduos da análise de regressão ordinária que é usada para estimar o desvio padrão sobre a linha de regressão

Histórico de comandos no R

```
bdreglog <- carregamento do dataset “Cap5_Exemplo_csv.csv” (dataset do livro do Corrar)
summary(bdreglog)
modrl=glm(ST ~ R+ND+VE, data = bdreglog, family = binomial)
summary(modrl)
prev.bdreglog <- predict(modrl, bdreglog, type="response")

plot(prev.bdreglog,bdreglog$ST)

newdata=data.frame(Rnovo=seq(min(bdreglog$R), max(bdreglog$R), len=92))
newdata$STpred=predict(modrl, bdreglog, type="response")
str(newdata)
plot(sort(STpred) ~ Rnovo, newdata)
```

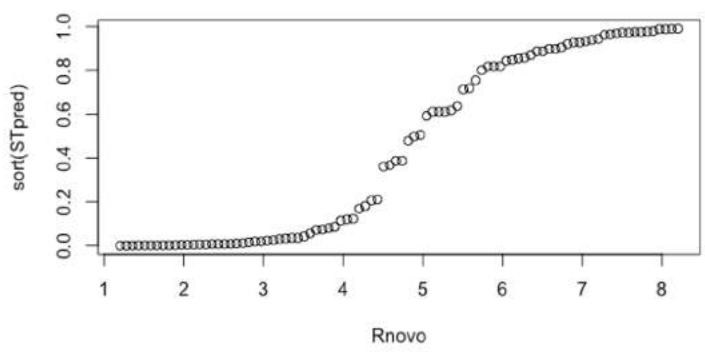
Construção do modelo de regressão logística

STpredito

Plot de ST X STpredito

Rnovo é uma sequência ordenada de 92 pontos equidistantes entre si criada no intervalo [Rmin,Rmax]

Plot de STpredito x Rnovo



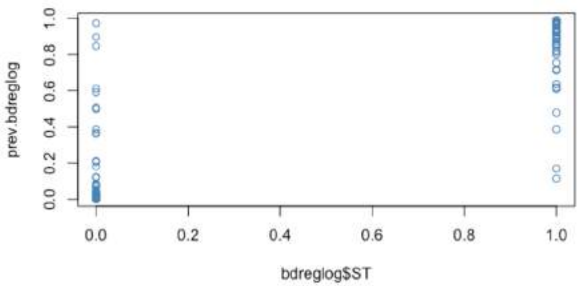
Histórico de comandos no R

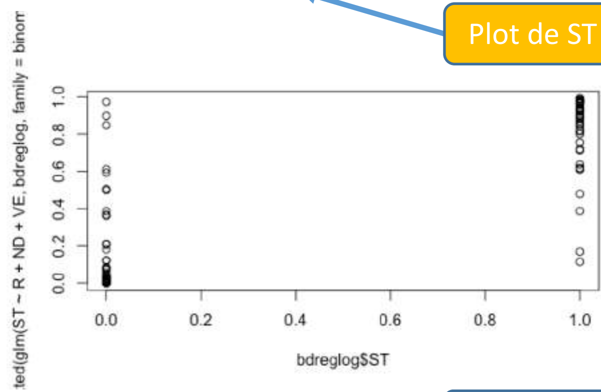
```
plot(bdreglog$ST,prev.bdreglog, col="steelblue")
```

Plot de ST x ST predito com modelo completo

```
plot(bdreglog$ST,fitted(glm(ST~R+ND+VE,bdreglog,family=binomial)))
```

Plot de ST x ST predito com modelo completo



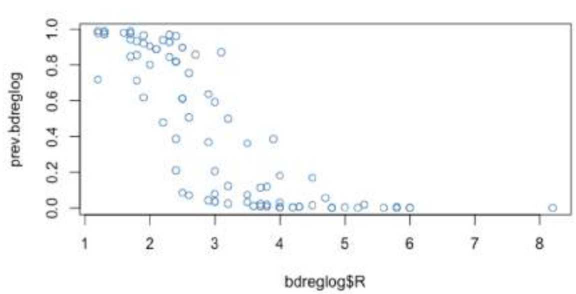


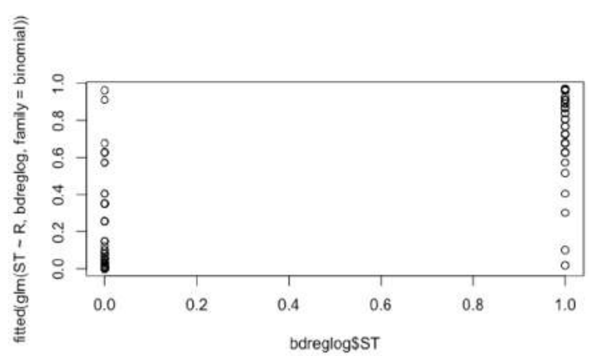

```
plot(bdreglog$R,prev.bdreglog, col="steelblue")
```

Plot de R x ST predito com modelo completo

```
plot(bdreglog$ST,fitted(glm(ST~R,bdreglog,family=binomial)))
```

Plot de ST x ST predito só com R

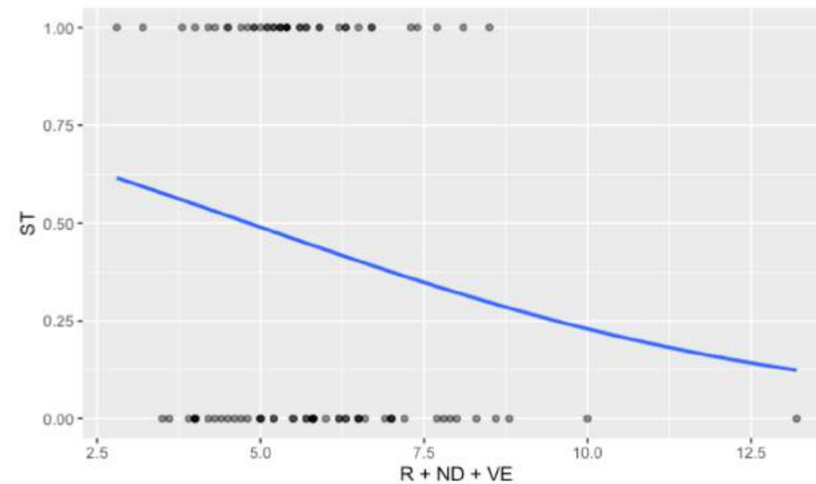




Histórico de comandos no R

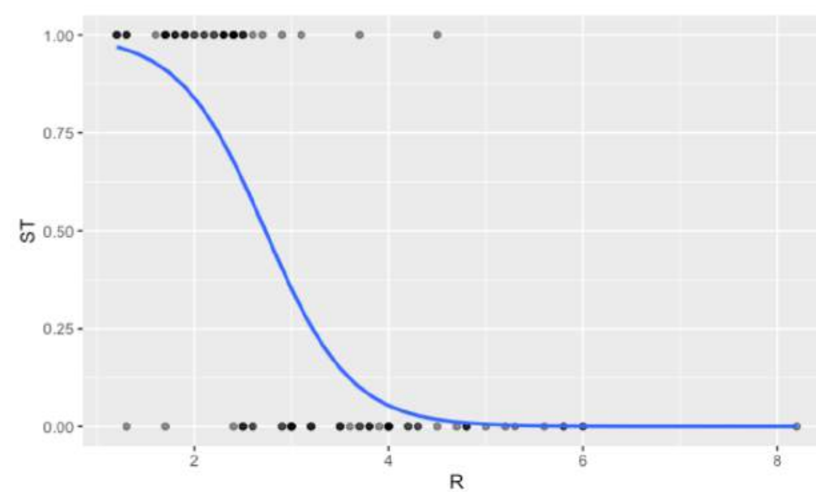
```
ggplot(bdreglog, aes(x=R+ND+VE, y=ST)) +
+ geom_point(alpha=.5) +
+ stat_smooth(method="glm", se=FALSE,
method.args = list(family=binomial))
```

Curva ST x R+ND+VE




```
ggplot(bdreglog, aes(x=R, y=ST)) +
+ geom_point(alpha=.5) +
+ stat_smooth(method="glm", se=FALSE,
method.args = list(family=binomial))
```

Curva ST x R



Histórico de comandos no R

Acurácia da predição:

Matriz de confusão

```
matconf <- table(bdreglog$ST,round(predict(modrl,bdreglog,type="response")))
      0  1  <- previsto
  0  45  6
  1   4 37
```

Acurácia

```
sum(diag(matconf))/sum(matconf)
[1] 0.8913043  <- 89,13%
```

Avaliação do modelo:

- R^2_{Logit} = -2(llhNull)-(-2llh)/-2(llNull) = (-2*(-63,225)-(-2*(-25,1537)))/-2*(-63,225) = 2407,057869
- Para avaliar os modelos use também o resultado McFadden
- Install (pscl) para usar a função pR2 (calcula um “pseudo-R2” para avaliação)

```
pR2(modrl)
```

fitting null model for pseudo-r2

llh	llhNull	G2	McFadden	r2ML	r2CU
-25.1536805	-63.2249871	76.1426132	0.6021560	0.5629192	0.7535501

Referências adicionais

- Métodos quantitativos em medicina, Análise multivariada, Prof. Raymundo Azevedo, https://www.youtube.com/watch?v=ou1Q90sUbNA&t=19s
- Evaluating Logistic Regression Models, Posted on August 17, 2015 by atmathew in R bloggers | 0 Comments, https://www.r-bloggers.com/2015/08/evaluating-logistic-regression-models/

Créditos: Profa. Ana Amélia Benedito Silva (conteúdo parcial de slides) *eof*