

# Regressão Logística

ACH2036 – Métodos Quantitativos Aplicados à Adm. de Empresas I

Prof. Regis Rossi A. Faria

2º sem. 2020



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

# Programa

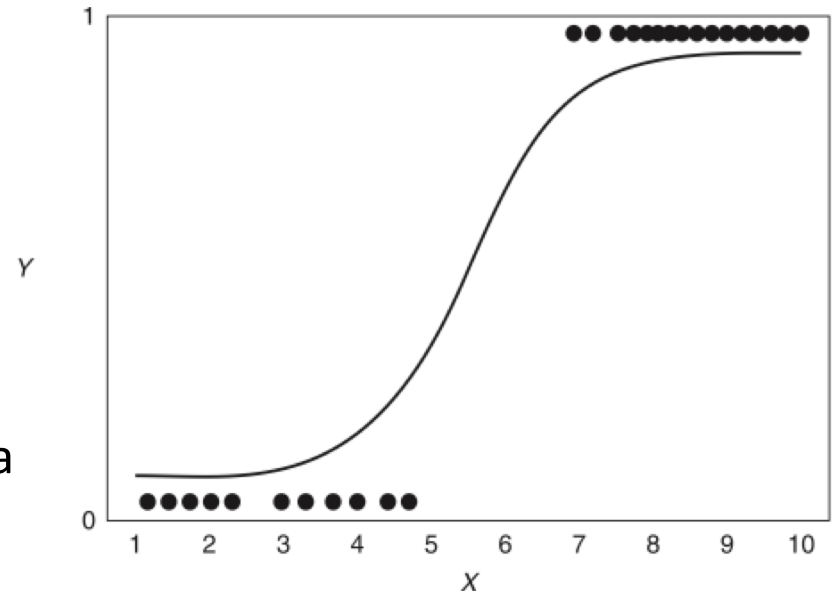
- Introdução (histórico, aplicabilidade)
- Modelo (equações usadas, propriedades)
- Exemplo
- Características (vantagens, suposições requeridas)

# 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)

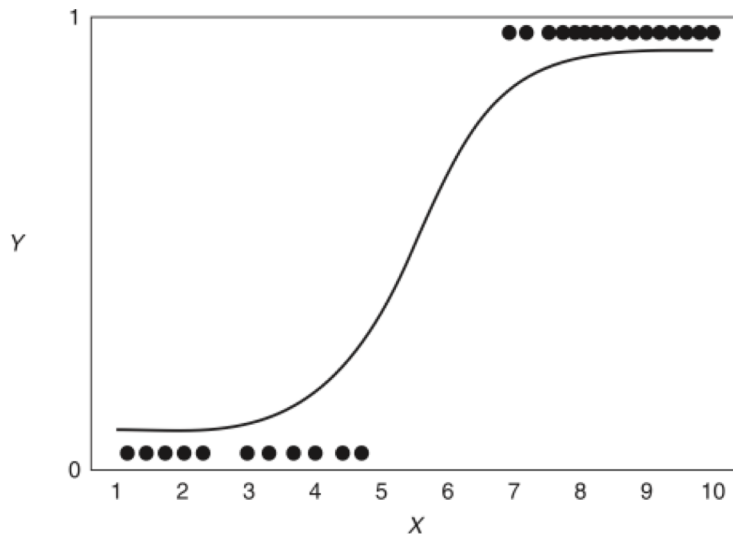
# Introdução

- 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 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

- O modelo que relaciona as variáveis independentes  $x_1, x_2, \dots$  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)

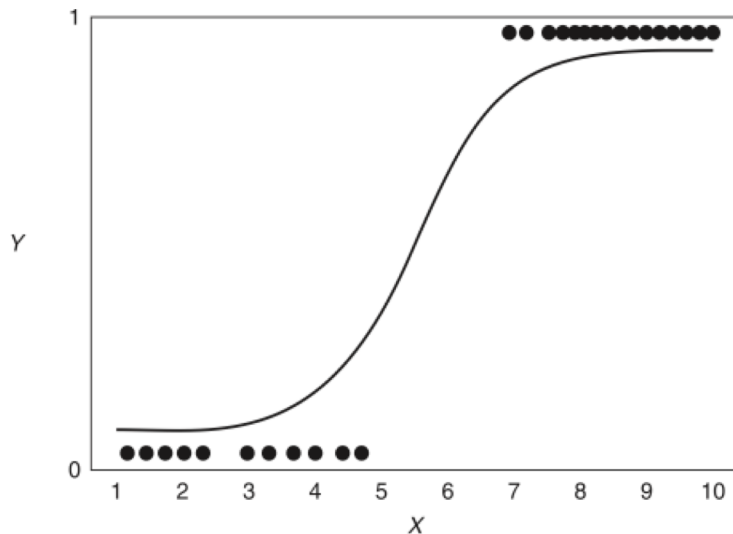


razão de chances

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

# Introdução

- O modelo que relaciona as variáveis independentes  $x_1, x_2, \dots$  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

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

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

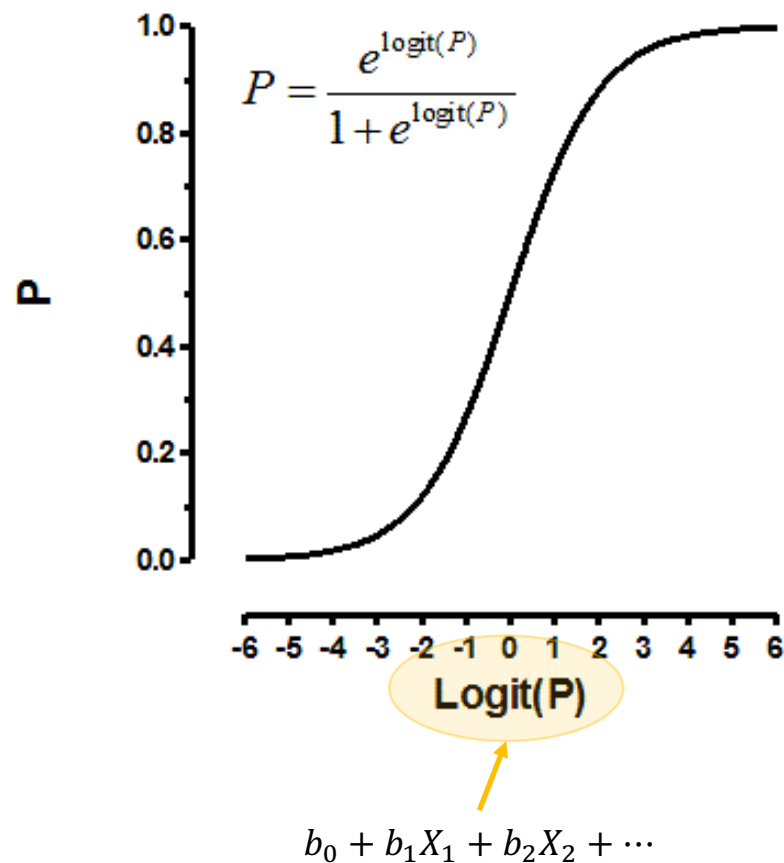
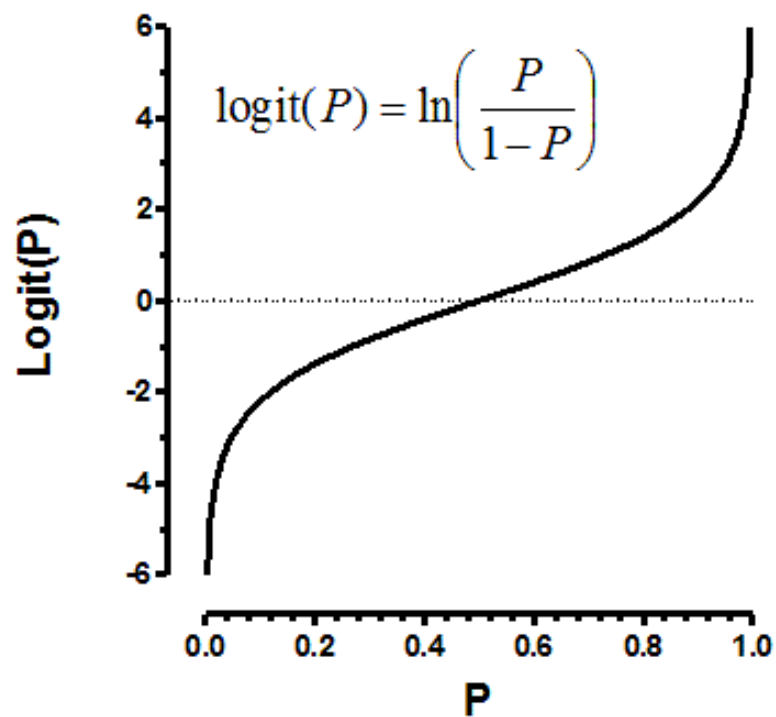
$$\text{logit}(Y) = \ln \left( \frac{Y}{1-Y} \right) = b_0 + b_1X_1 + b_2X_2 + \dots$$

agora com  $Y = P(\text{evento})$  pode ser reescrita equivalentemente como

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

razão de chances

# P(Y) e logit





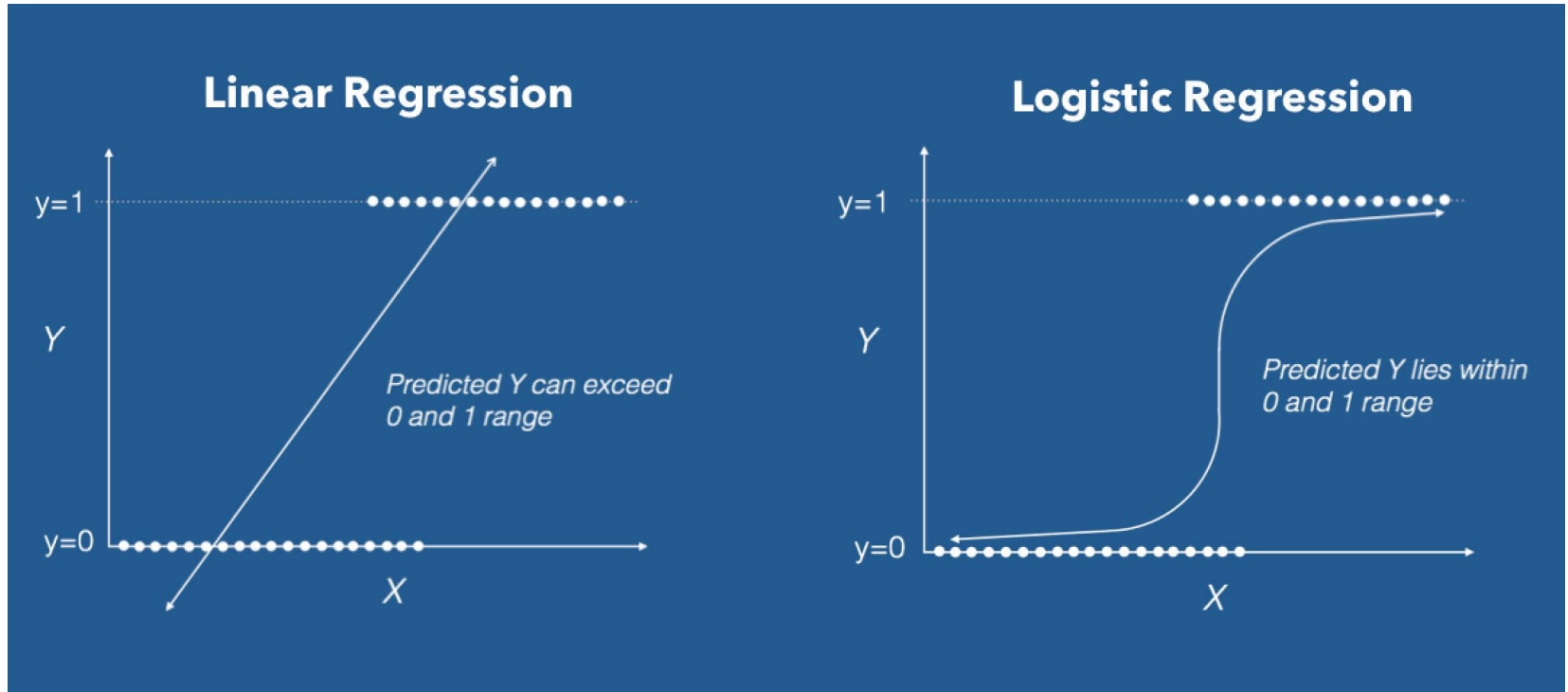
# Valores comparados

- Probabilidades variam de 0 a 1
- Razão de chances varia de 0 a  $+\infty$  (NC) (crescente)
- *logit* varia numa faixa entre  $-\infty$  (NC) e  $+\infty$  (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



# 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  $\mathbf{X}$ , a técnica prevê uma probabilidade  $0 < Y < 1$ , usando os coeficientes  $\mathbf{b}$  estimados
  - ✓ Se  $Y > 0,5 \rightarrow Y = 1$
  - ✓ Se  $Y \leq 0,5 \rightarrow Y = 0$

# Regressão Logística

Modelos de regressão não linear são usados, em geral, 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). 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.

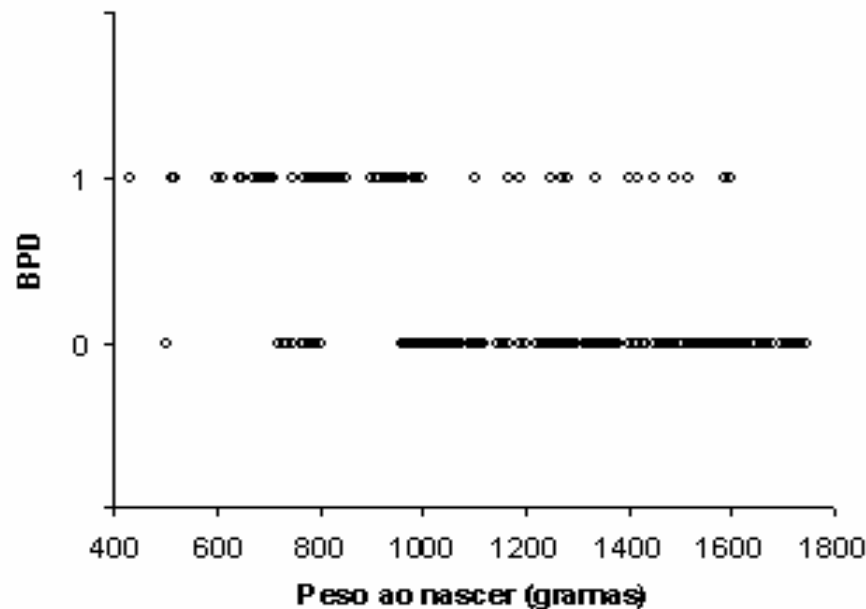
## Exemplo:

Bebês, ao nascer, abaixo de 1750 gramas estão confinados em uma UTI neonatal. Em uma amostra de 223 bebês, 76 apresentaram diagnóstico com displasia broncopulmonar (BPD).

A probabilidade de uma criança, nessas condições ter BPD é

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

**Análise gráfica:**



# Modelo Geral

## ***Conceitos***

Modelo de Regressão Múltipla

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

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

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



## 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:

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

## MODELO GERAL

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

## FUNÇÃO LOGÍSTICA

Modelo inicial:

$$p = \beta_0 + \beta_1 x$$

sendo  $x$  o peso ao nascer. Para que  $0 < p < 1$ , então o modelo é dado por

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

## MODELO GERAL

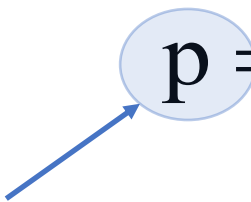
$$E(Y_i) = \beta_0 + \beta_1 X_i$$

## FUNÇÃO LOGÍSTICA

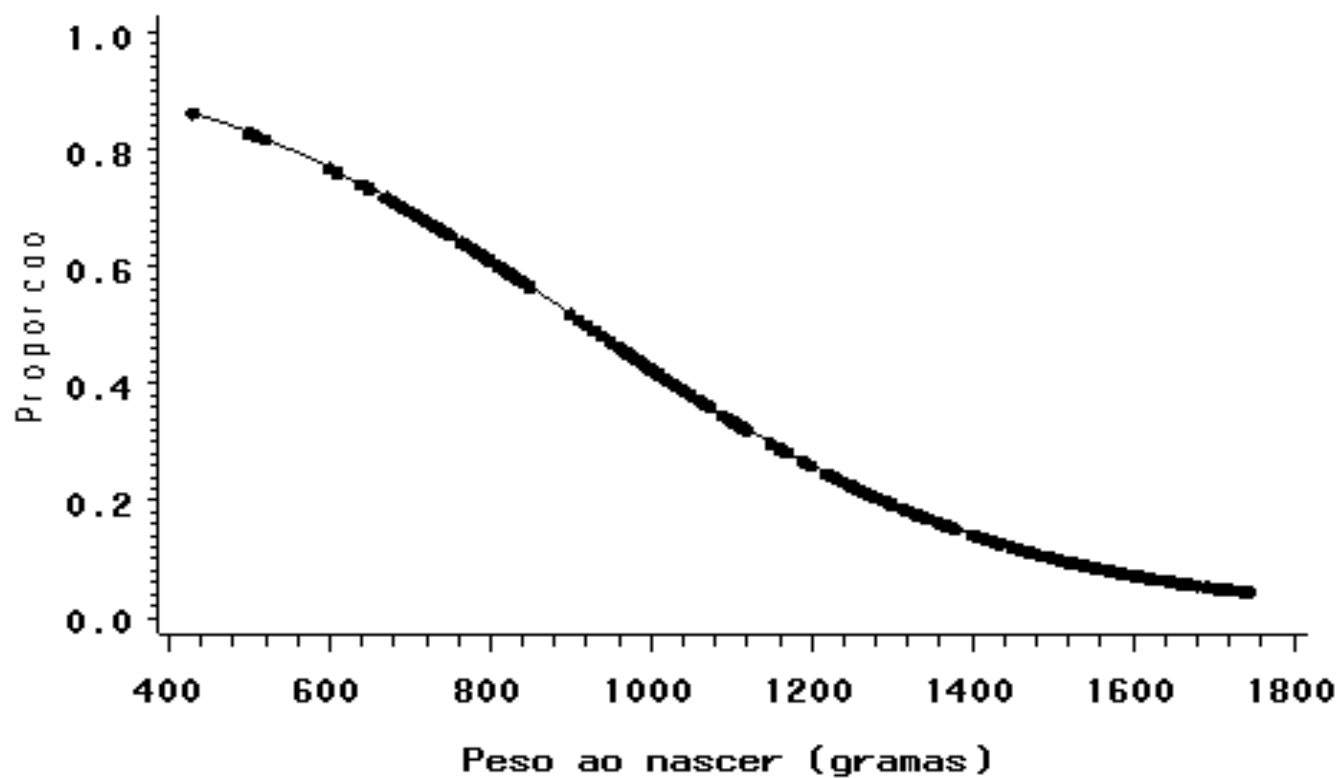
Modelo inicial:

$$p = \beta_0 + \beta_1 x$$

sendo  $x$  o peso ao nascer. Para que  $0 < p < 1$ , então o modelo é dado por


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

nova variável (transformada)



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

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

**Fator:** o peso de nascimento do bebê (0 |-- 950, 950 |-- 1350, 1350 |-- 1750)

**Variável resposta:** o bebê está ou não está doente

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

## Modelo para o conjunto de 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

**Através desta função posso afirmar que o peso está relacionado com a presença de BPD?**

# 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

$$e^{b_1}, \dots, e^{b_k}$$

como uma razão de chances  
(*odds ratio*)

# Regressão logística

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

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

Se:

$e^{b_1} = 1$ , então a chance de  $x_1$  apresentar  $y=1$  é a mesma que  $x_3$

$e^{b_1} > 1$ , então a chance de  $x_1$  apresentar  $y=1$  é maior que  $x_3$

$e^{b_1} < 1$ , então a chance de  $x_1$  apresentar  $y=1$  é menor que  $x_3$



## Analizando a relação entre duas variáveis

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 <sup>a</sup>	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.

O que significa a primeira linha da tabela, referente a PESO ?

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

## Analizando a razão de chances (OR)

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

**Veja no youtube:**

<https://www.youtube.com/watch?v=ou1Q90sUbNA&t=19s>

# 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-quadrado do modelo logístico
  - Teste Cox-Snell  $R^2$
- Testes usados:
  - Hosmer e Lemeshow
  - Teste Wald

# 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: 0 (ajuste perfeito)
  - Serve para verificar se o modelo melhora com a inclusão/exclusão de alguma variável independente
- R-quadrado do modelo logístico
  - pseudo-R<sup>2</sup>
  - R<sup>2</sup>logit 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)
  - 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: distribuição qui-quadrado
    - $Wald = (b/SE)^2$

# 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: não requer normalidade nem variância constante (homoscedasticidade). No entanto requer que o valor esperado do erro seja 0; inexistência de autocorrelação entre erros; e entre estes e as variáveis independentes; e ausência de multicolinearidade perfeita entre as variáveis independentes
- Atraente para aplicações de *machine learning*
- Facilidade em prever 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

- $\text{logit}(p) = \ln(p/(1-p))$
- $\ln(\text{odds}) = \ln(p/(1-p))$
- $\text{odds} = p / (1-p)$
- Na regressão logística não assumimos uma relação linear entre a variável dependente e independente
- Erros não têm distribuição normal
- Utilizamos a máxima verossimilhança, e não mínimos quadrados

# Exemplo

- 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

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

Legenda:

ST = status de adimplência

R = renda mensal

ND = número de dependentes

VE = vínculo empregatício

- Realizando previsões com o modelo:

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



# Interpretando o impacto de uma variável

- Exemplo:  $\text{logit} = 0,25x_1 + 0,4x_2$
- $x_1$  = 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  $\rightarrow 1,49*0,43=0,64$
- logo  $p$  passou para  $p'=0,39$  ( $p \sim 0,4$ )  $\rightarrow$  um aumento de  $\sim 10\%$

FIM