

Agenda

Aula 3



01

R e g r e s s ã o L i n e a r e

M ú l t i p l a

02

R e g r e s s ã o L o g i s t i c a

03

Á r v o r e s d e D e c i s ã o

2.3 Sessão Síncrona nº 3 – 9 de Fevereiro de 2021 (21h00-23h00)

Regressão Linear

- Regressão Linear - Simples
- Regressão Linear - múltipla
- Regressão Não Linear - Logística

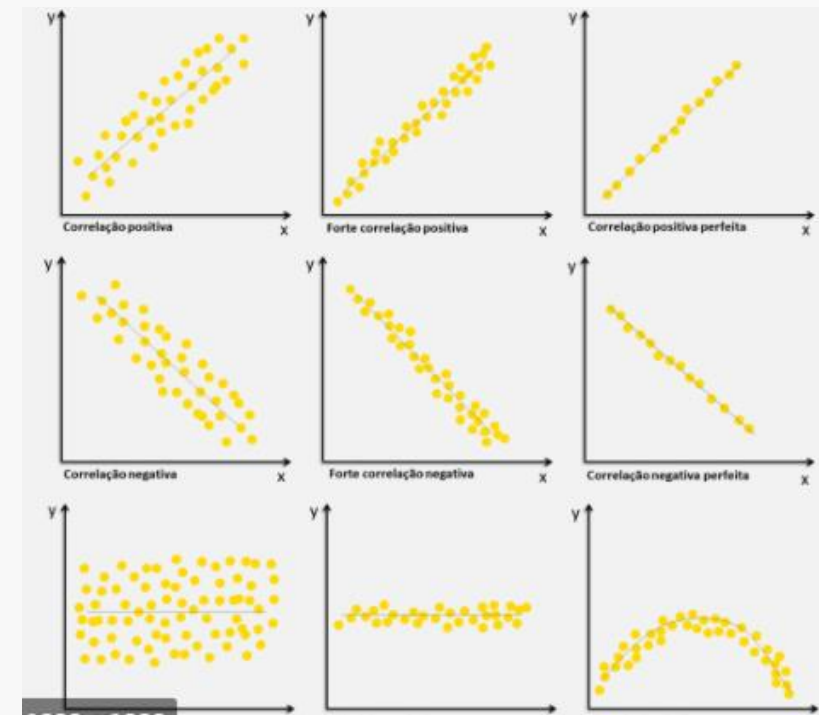
Breve abordagem às Árvores de decisão

- Conceito
- Representação
- Características básicas
- Cálculo de entropia

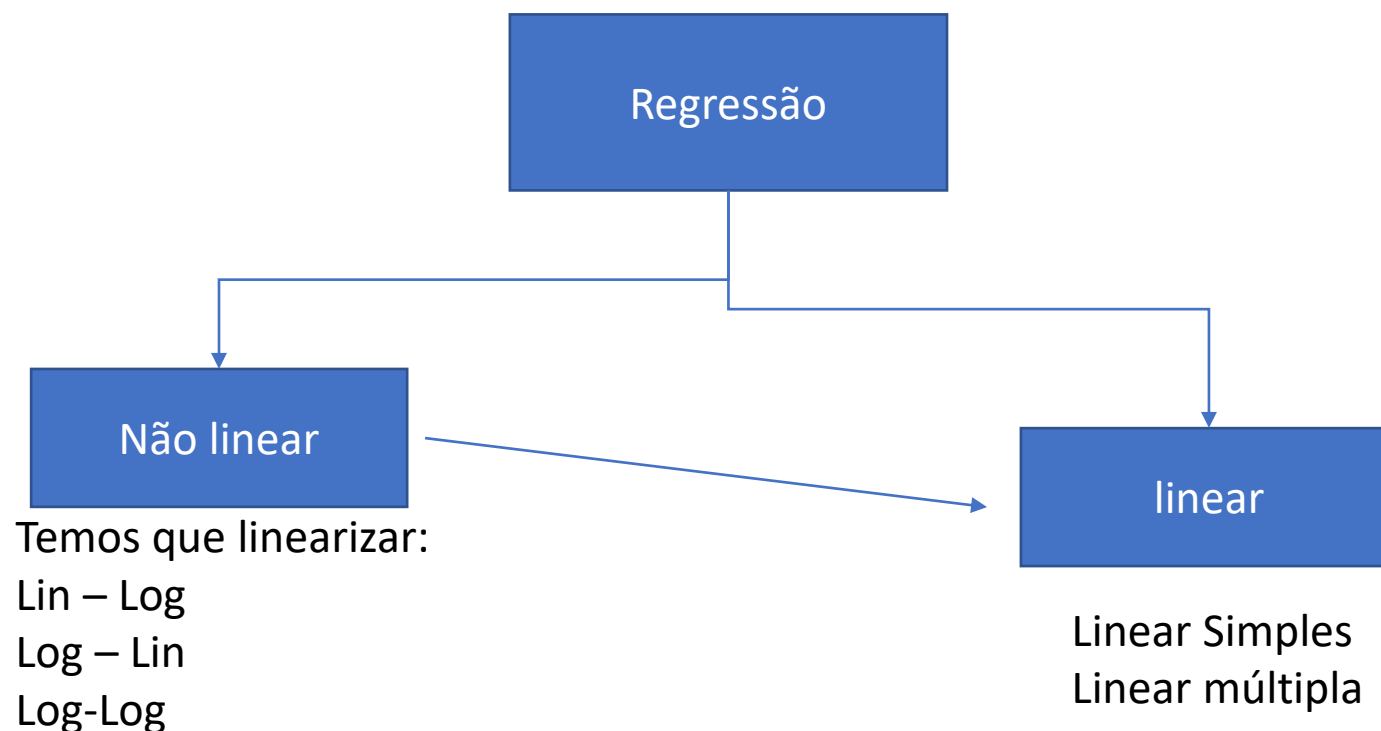


01

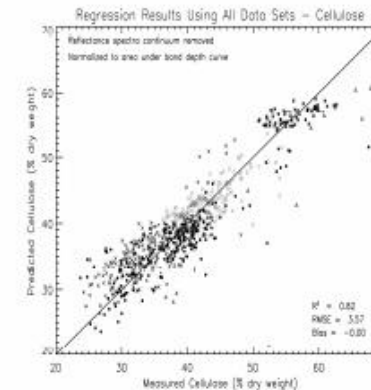
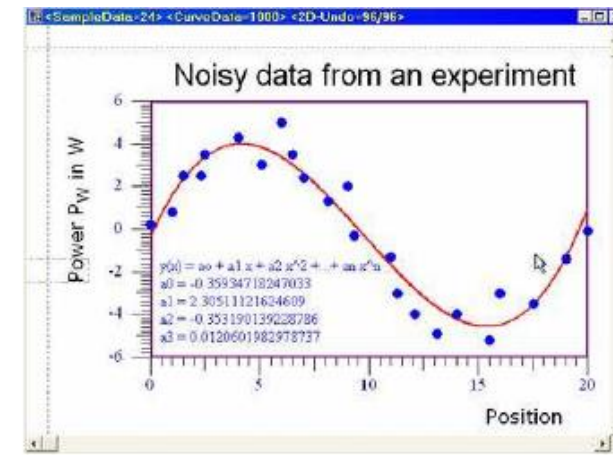
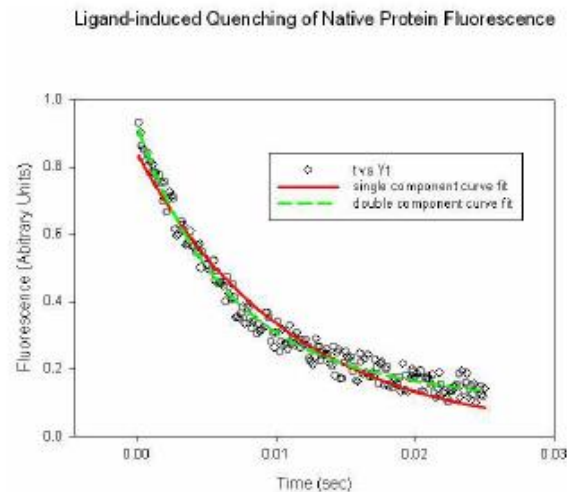
Regressão Linear Simples



Relação entre variáveis

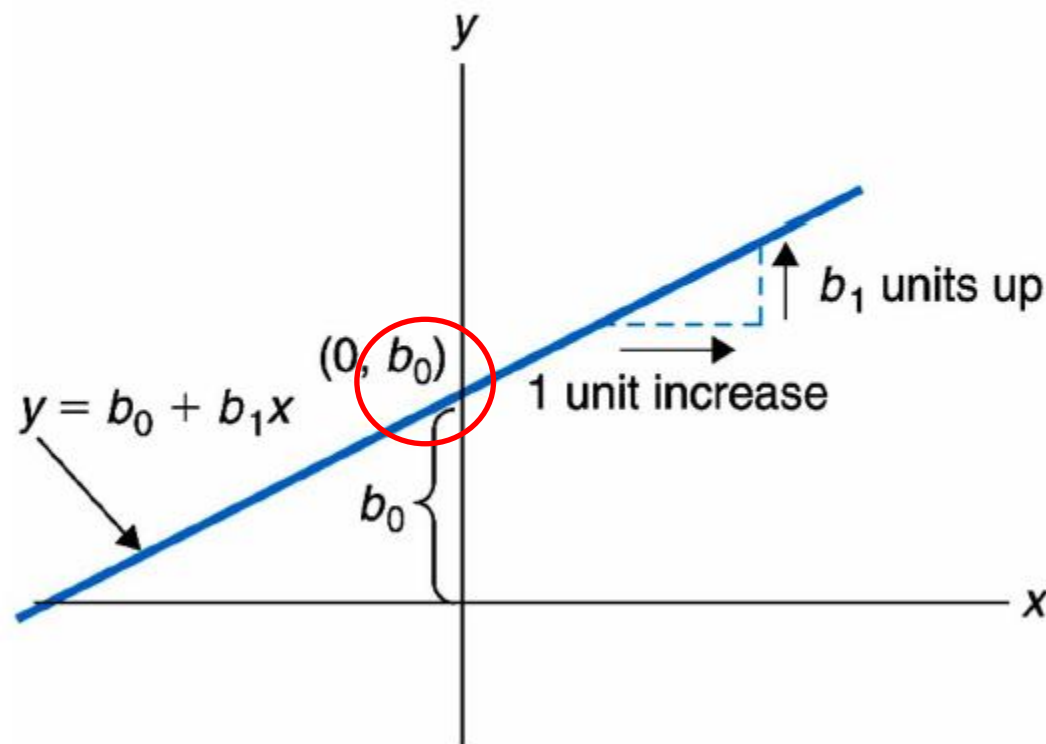


Relação entre duas variáveis (dois eixos de análise)



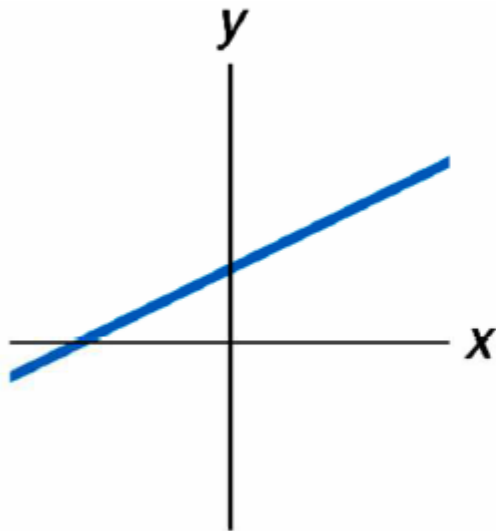
Modelo de Regressão Linear

Para uma equação de uma recta $y = b_0 + b_1x$, ao valor b_0 chama-se **ordenada na origem** e ao valor b_1 chama-se **declive**.

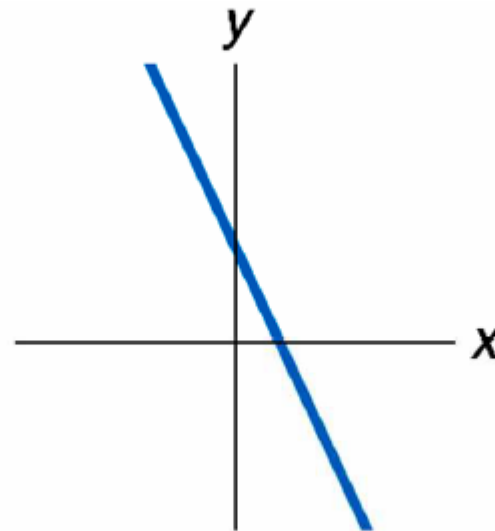


Modelo de Regressão Linear

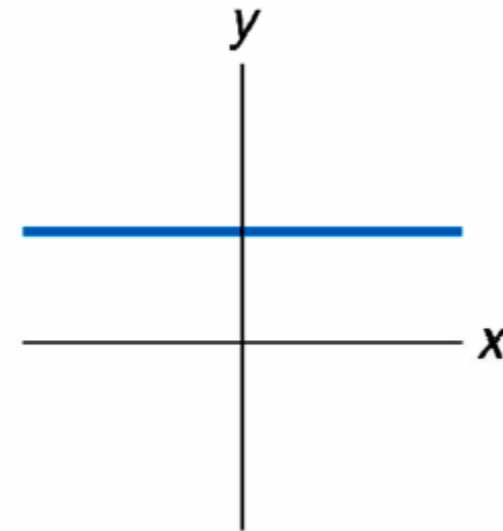
Interpretação gráfica do declive:



$$b_1 > 0$$



$$b_1 < 0$$



$$b_1 = 0$$

Modelo de Regressão Linear

Pretende estabelecer uma função matemática que descreva a relação entre uma variável contínua (variável explicada ou dependente) e uma ou mais variáveis explicativas ou independentes.

$$y = f(x_1, x_2, \dots, x_k) + \varepsilon$$

y denota a variável dependente;

x_1, x_2, \dots, x_k denotam as variáveis independentes;

$f(x_1, x_2, \dots, x_k)$ descreve a variação sistemática;

ε representa a variação não sistemática (erro aleatório).

Modelo de Regressão Linear

O objetivo da análise de regressão linear consiste em identificar uma equação linear que permita prever o valor da variável dependente em função dos valores conhecidos das variáveis independentes.

Regressão linear simples: apenas uma variável independente.

Exemplo:

- *variável dependente = vendas*
- *variável independente = despesas com marketing*

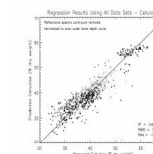
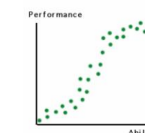
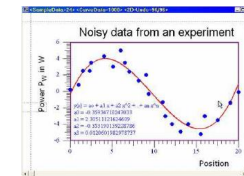
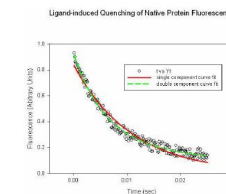
Regressão linear múltipla: duas ou mais variáveis independentes.

Exemplo:

- *variável dependente = preço do imóvel*
- *variáveis independentes = área, nº de quartos, nº de WC, idade*

Diagrama de Dispersão

- Um diagrama de dispersão mostra a relação entre duas variáveis quantitativas, medidas sobre a mesma observação.
- Os valores de uma variável aparecem no eixo horizontal, e os da outra, no eixo vertical.
- Comumente coloca-se no eixo x a variável independente.
- Cada indivíduo aparece como o ponto do gráfico definido pelos valores de ambas as variáveis para determinada observação (x,y).



Exemplos de relações entre variáveis

Produção

Número de peças produzidas e número de peças defeituosas

Construção

Número de falhas em uma obra e a satisfação média dos construtores

Dias de atraso de entrega x número de dias chuvosos

Área financeira

Média de tempo de atraso de pagamento e número de erros de fatura

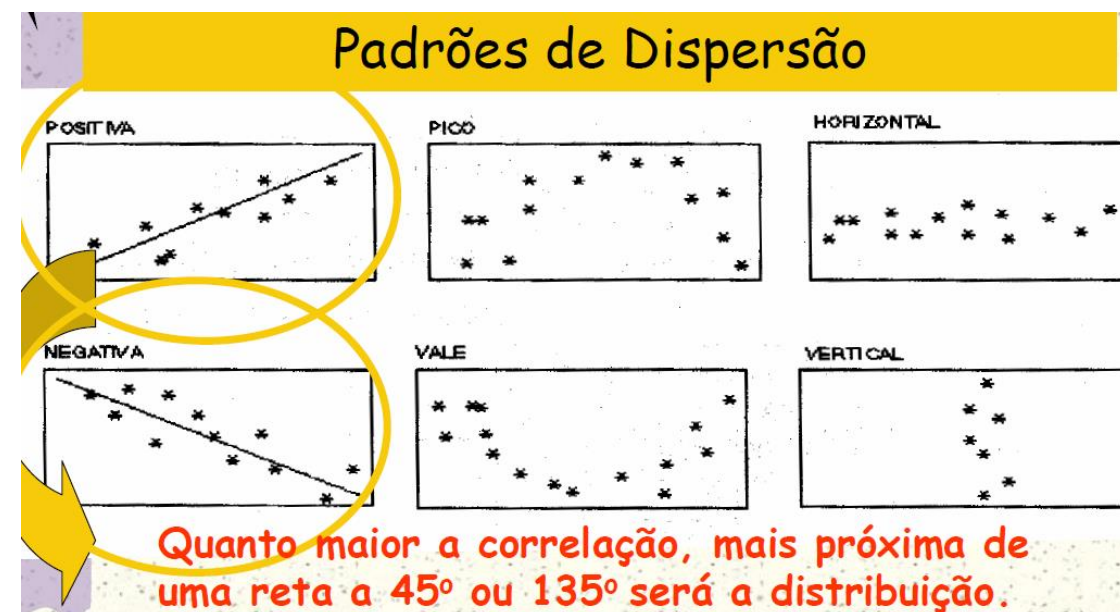
Vendas

% de imóveis vendidos na data de entrega da obra x satisfação média dos clientes nos últimos 10 empreendimentos.

Aspectos a ter em conta na análise

Os aspectos abaixo são relevantes na análise dos diagramas:

- DIREÇÃO (crescente, decrescente);
- FORMA (linear, não-linear, aglomerados);
- PONTOS DISCREPANTES.



Duas tipologias de Modelo Regressão Linear

Modelo de regressão linear simples:

- uma variável dependente explicada por uma variável independente.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Modelo de regressão linear múltipla:

- Uma variável dependente explicada pelo menos por duas variáveis independentes.

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_K x_K + \varepsilon \quad (K \geq 2)$$

Explicação dos Erros

O erro (ε) representa:

- Todos os outros fatores que afetam a variável dependente Y , mas que não estão contempladas nas variáveis explicativas X_1 , X_2 , etc...

Algumas razões para a existência dos erros:

- Erros de medição.
- Forma funcional inadequada, por exemplo, $y = \beta_0 + \beta_1 x$ ou $y = \beta_0 + \beta_1 x + \beta_1 x^2$?
- Inerente variabilidade no comportamento dos agentes económicos.

Modelo Regressão Linear

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i=1, n$$

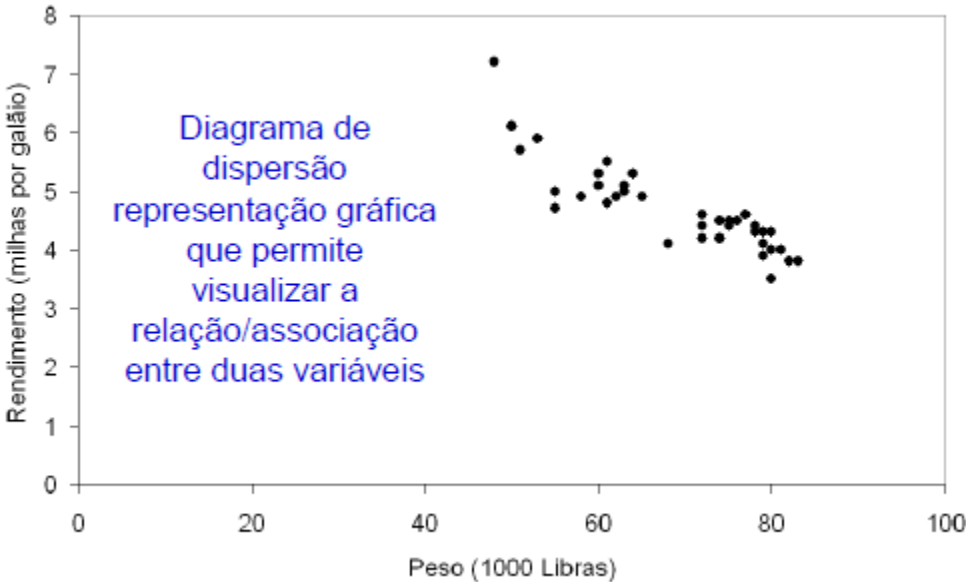
Componente
determinística

Erro, uma variável aleatória
não-observável

Exemplo

Consumo de combustível de camiões de acordo com a carga transportada

peso	Milhas por galão
60	5.3
55	5
80	4
72	4.2
75	4.5
63	5.1
48	7.2
79	3.9
82	3.8
72	4.4
58	4.9
60	5.1
74	4.5
80	4.3
53	5.9
61	5.5
80	3.5
68	4.1
76	4.5
75	4.4
63	5
65	4.9
72	4.6
81	4
64	5.3
78	4.4
62	4.9
83	3.8
79	4.1
61	4.8
63	5
62	4.9
77	4.6
76	4.5
51	5.7
74	4.2
78	4.3
50	6.1
79	4.3
55	4.7

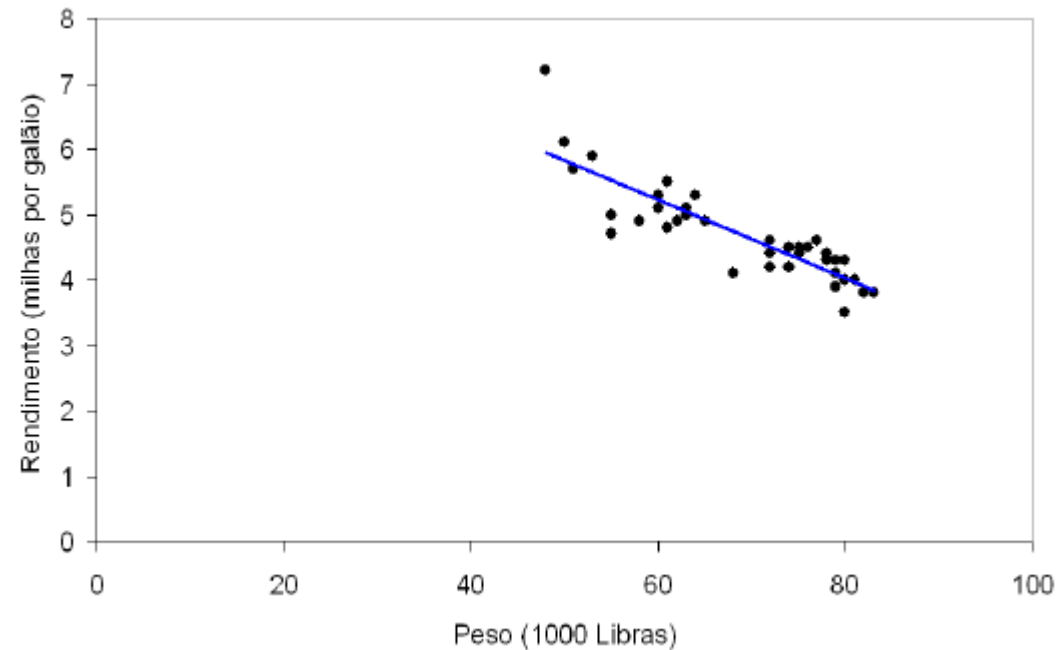


Um incremento no peso reduz o rendimento

A relação entre as variáveis não é exata (estocástica)

Exemplo

Consumo de combustível de caminhões de acordo com a carga transportada



modelo

$$y = \beta_0 + \beta_1 x + \varepsilon$$

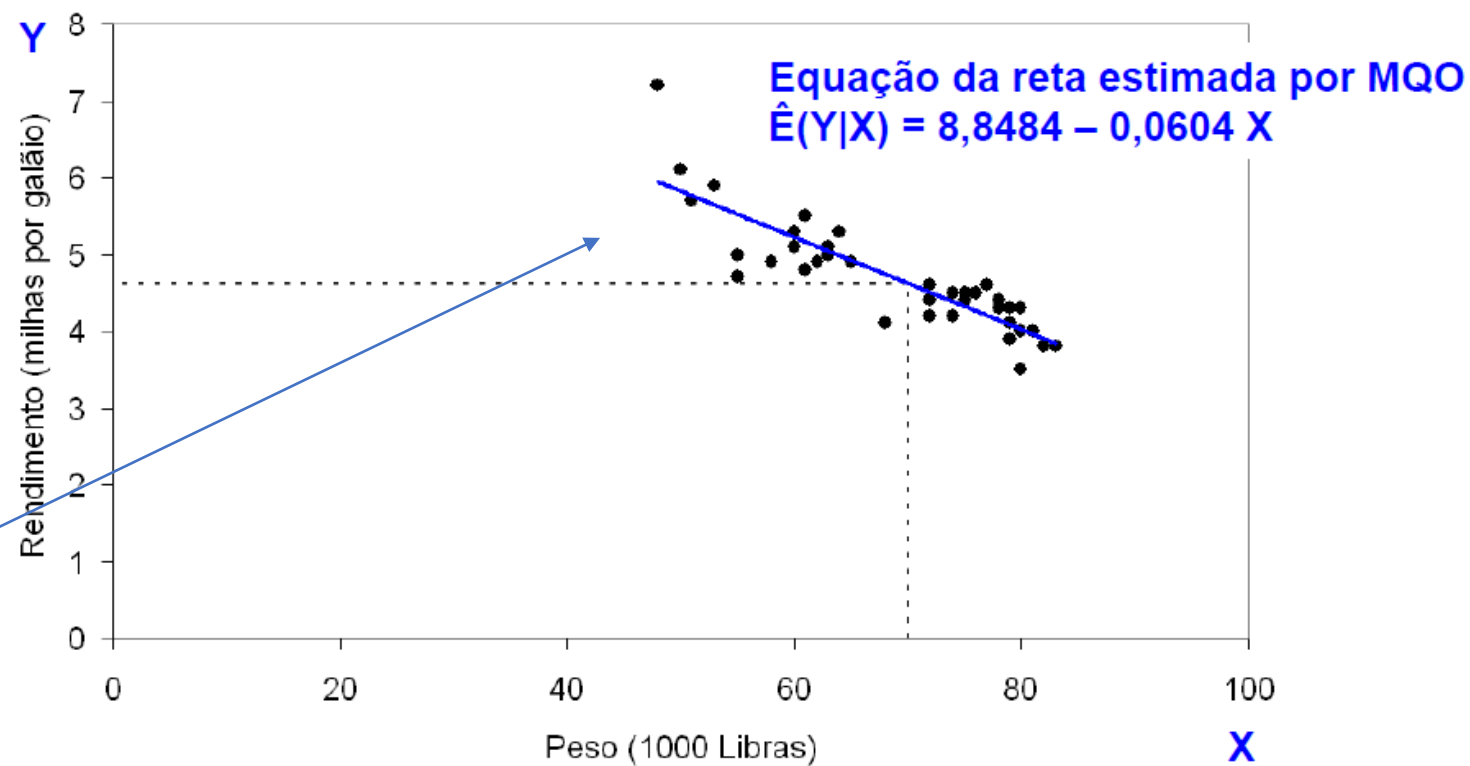
β_0 e β_1 são constantes não conhecidas

ε é um termo aleatório com distribuição normal ($\varepsilon \sim N(0, \sigma^2)$)

Exemplo

Consumo de combustível de caminhões de acordo com a carga transportada

Resultado da estimação da Reta



Exemplo

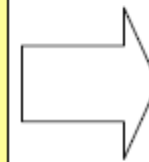
Consumo de
combustível de caminhões
de acordo com a carga
transportada

Estimação de parâmetros

Estimador MQO

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Modelo ajustado

$$\hat{E}(Y|X) = 8,8484 - 0,0604 X$$

$\hat{\beta}_0$

$\hat{\beta}_1$

Hipóteses assumidas pelo modelo

H1) A relação entre as variáveis é linear $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i=1, n$:

H2) Média nula: $E(\varepsilon_i) = 0$ para todo $i=1, n$

H3) Variância constante: $V(\varepsilon_i) = \sigma^2$ para todo $i=1, n$

H4) Erros não correlacionados: $Cov(\varepsilon_i, \varepsilon_k) = 0$ para todo $i \neq k$

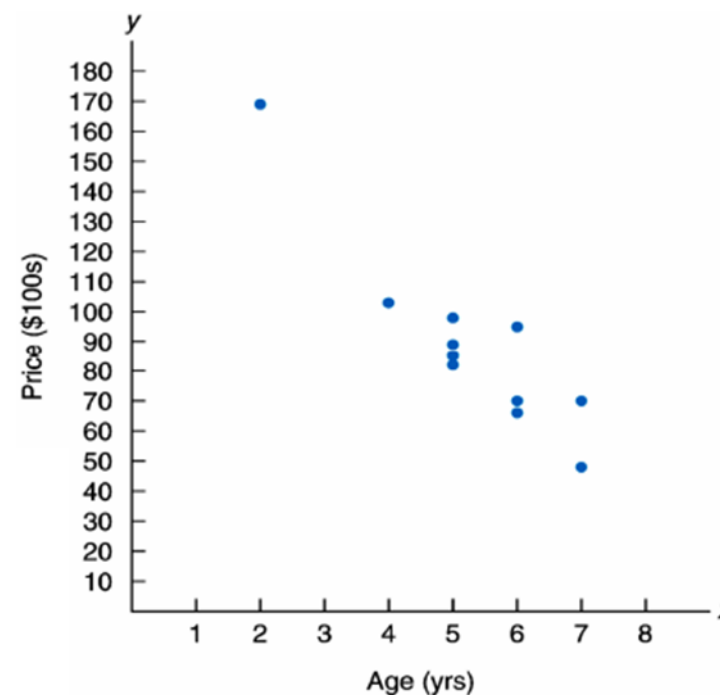
H5) Distribuição Normal: $\varepsilon_i \sim N(0, \sigma^2)$ para todo $i=1, n$

ε_i **são independentes e identicamente distribuídos $N(0, \sigma^2)$**

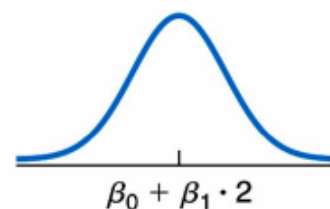
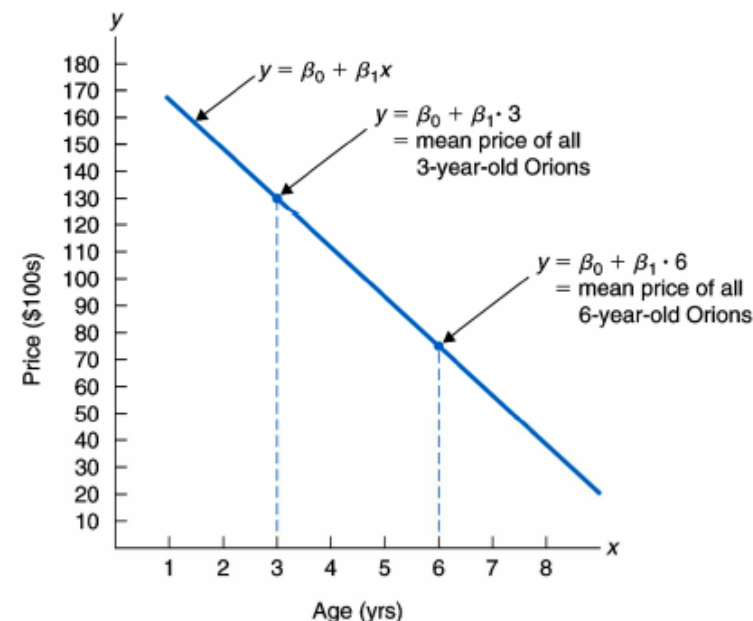
H6) A variável explicativa X é fixa, i.e., não é estocástica

Vejamos mais um caso: Preços de carros usados de acordo com a idade

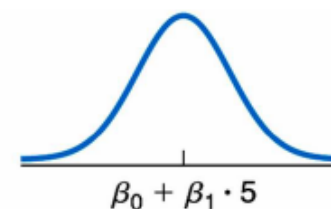
Car	Age (yrs) x	Price (\$100s) y
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48



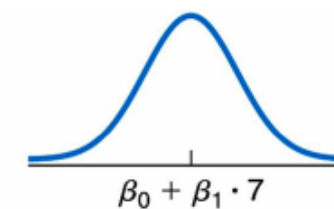
Vejamos mais um caso: Preços de carros usados de acordo com a idade



Prices of 2-year-old Orions



Prices of 5-year-old Orions



Prices of 7-year-old Orions

Parâmetros do modelo

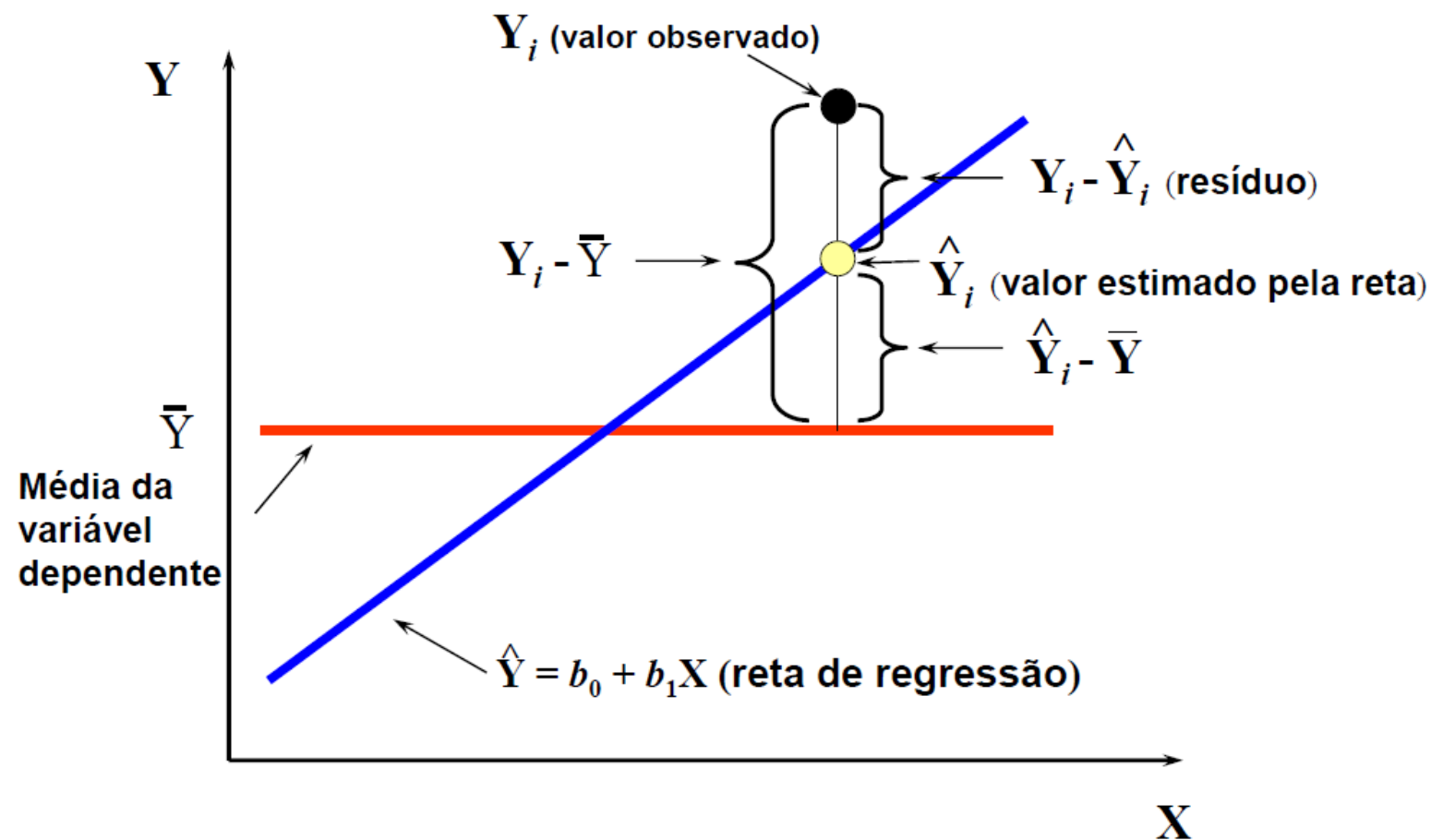
Valor estimado da variável dependente y dado que x é igual a x_i

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Resíduo da i -ésima observação é igual a diferença entre o valor observado e o valor estimado da variável y_i

$$\begin{aligned}\hat{\epsilon}_i &= y_i - \hat{y}_i \\ \hat{\epsilon}_i &= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\end{aligned}$$

Análise do Erro e sua decomposição



Análise do Erro e sua decomposição

Os estimadores
são normalmente
distribuídos

$$\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2)$$

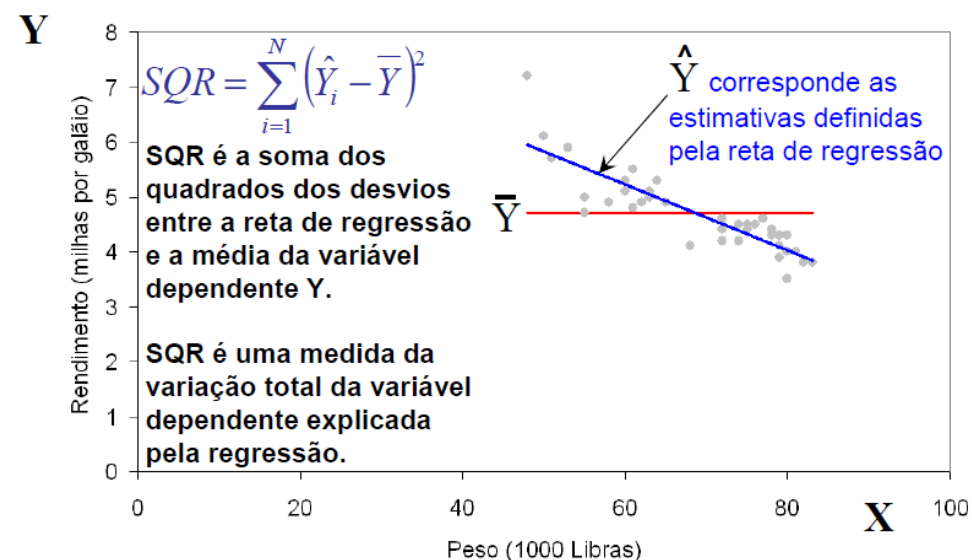
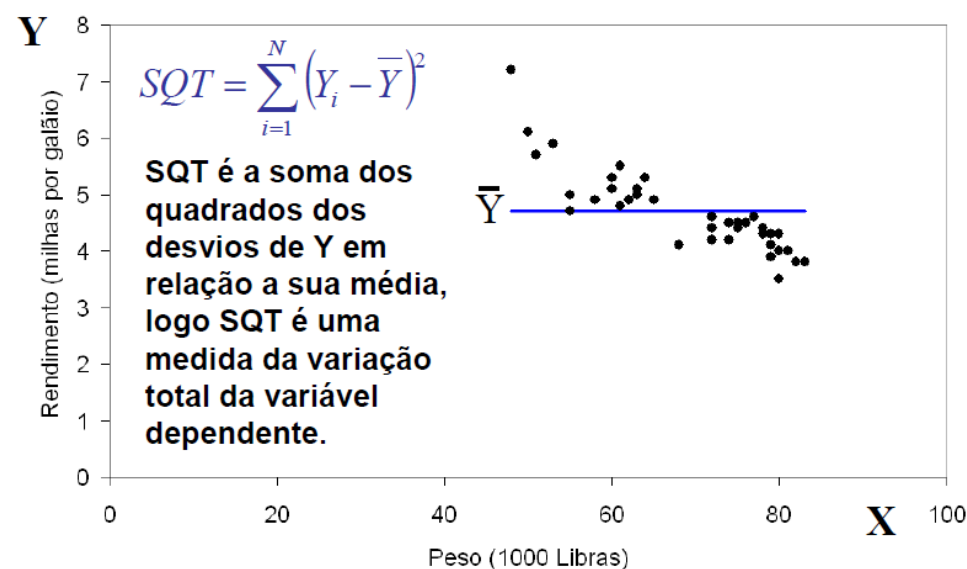
$$\hat{\beta}_1 \sim N(\beta_1, \sigma_{\hat{\beta}_1}^2)$$

Se as hipóteses H1 até H6 forem satisfeitas, os estimadores de mínimos quadrados são estimadores lineares não tendenciosos de variância mínima (Teorema de Gauss Markov)

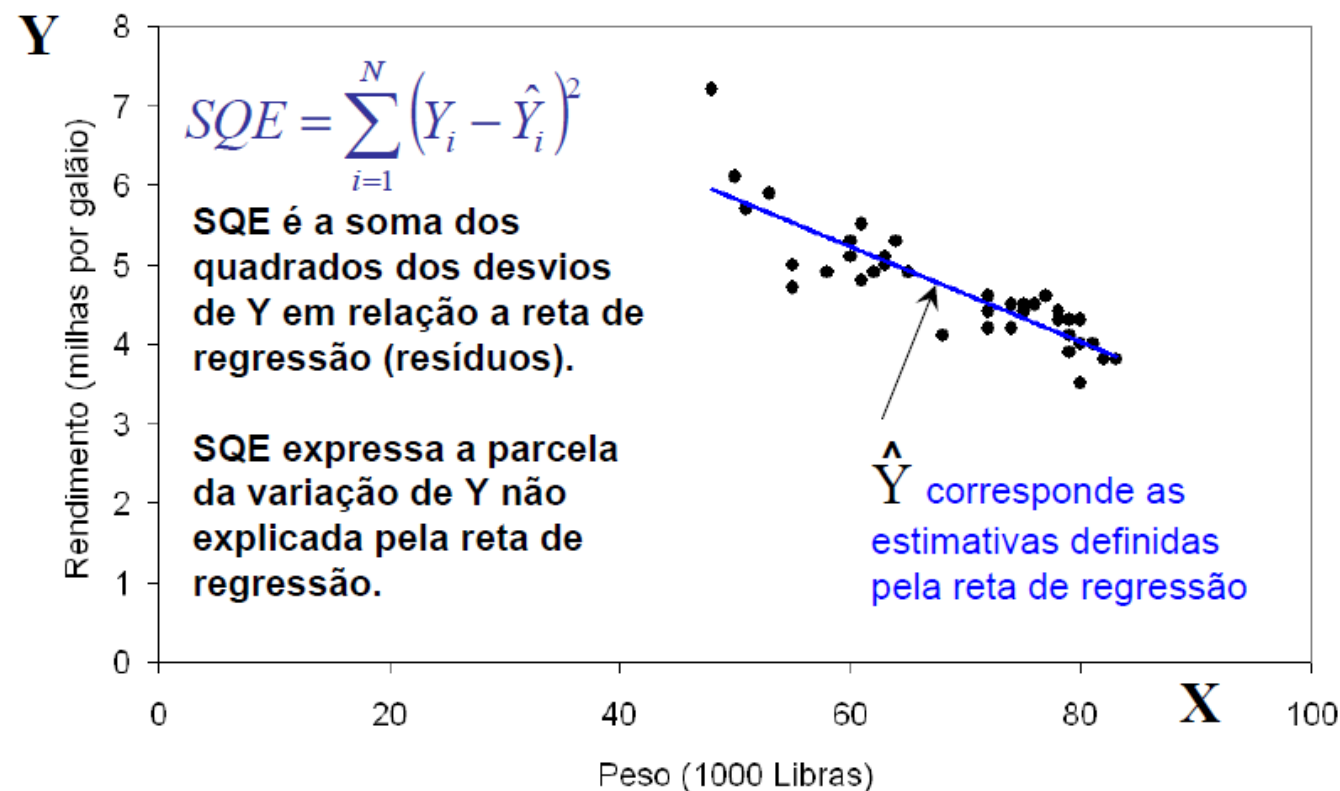
Estimador da
variância do
erro

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

Análise do Erro e sua decomposição



Análise do Erro e sua decomposição



Análise do Erro e sua decomposição

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2$$

$$\text{SQT} = \text{SQE} + \text{SQR}$$

SQT = Soma de Quadrados Total

SQR = Soma de Quadrados da Regressão

SQE = Soma de Quadrados dos Erros (Resíduos)

Coeficiente de determinação

O coeficiente de determinação, também chamado de R^2 , é uma medida de ajuste de um modelo estatístico linear generalizado, como a regressão linear simples ou múltipla, aos valores observados de uma variável aleatória.

O R^2 varia entre 0 e 1, por vezes sendo expresso em termos percentuais. Nesse caso, expressa a quantidade da variância dos dados que é explicada pelo modelo linear. Assim, quanto maior o R^2 , mais explicativo é o modelo linear, ou seja, melhor ele se ajusta à amostra.

Por exemplo, um $R^2 = 0,8234$ significa que o modelo linear explica 82,34% da variância da variável dependente a partir do regressores (variáveis independentes) incluídas naquele modelo linear.

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2} = 1 - \frac{SQE}{SQT}$$

$$0 \leq R^2 \leq 1$$

R² ajustado

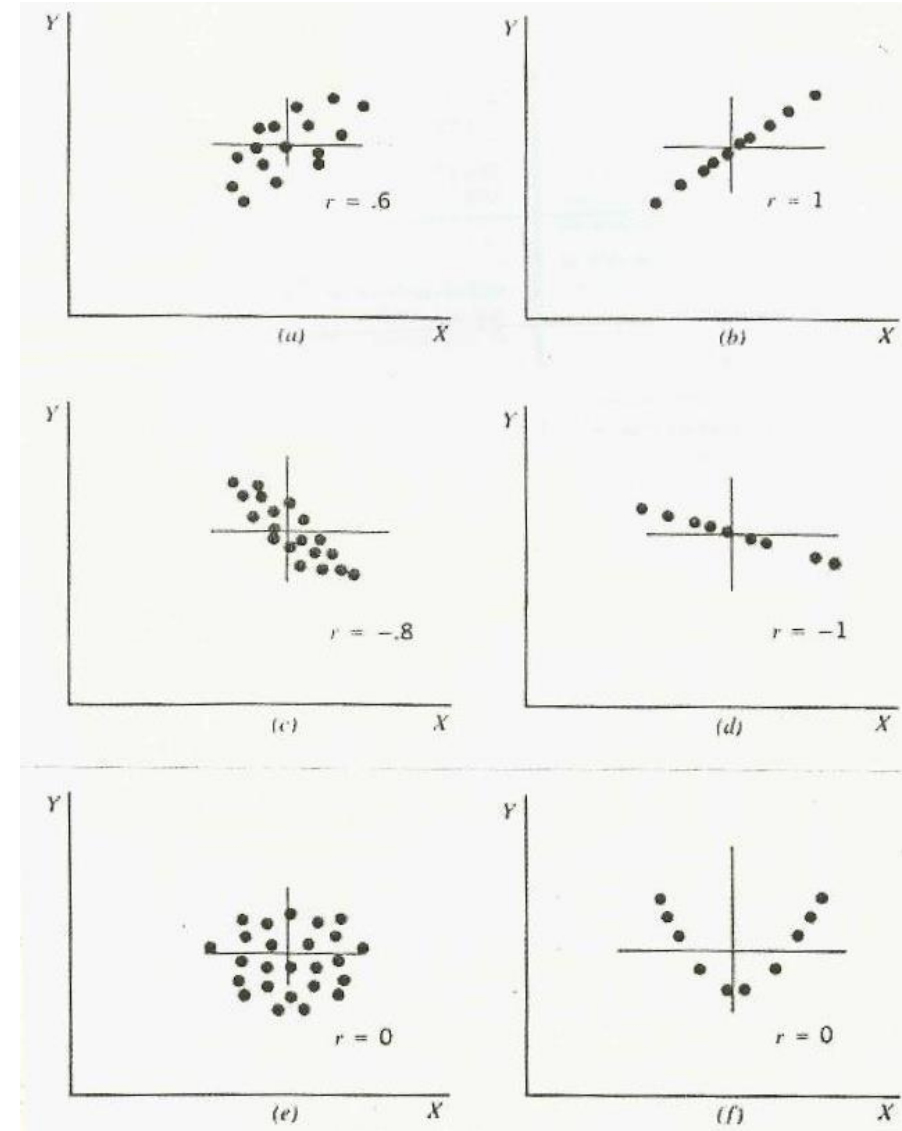
A inclusão de inúmeras variáveis, mesmo que tenham muito pouco poder explicativo sobre a variável dependente, aumentarão o valor de R². Isto incentiva a inclusão indiscriminada de variáveis, prejudicando o princípio da parcimônia (ver de forma mais ampla em navalha de Ockham).

Para combater esta tendência, podemos usar uma medida alternativa do coeficiente de determinação, que penaliza a inclusão de regressores pouco explicativos:

$$\bar{R}^2 = 1 - \frac{n-1}{n-(k+1)} (1 - R^2),$$

onde (k+1) representa o número de variáveis explicativas mais a constante.

Interpretação de R



Análise de resultados

Análise da variância (ANOVA)

Causas de variação	Graus de liberdade	Soma dos quadrados	Quadrados médios
Regressão	1	$SQR = \hat{\beta}_1^2 \sum_{i=1}^N (x_i - \bar{X})^2$	$QMR = SQR/1$
Resíduos	N - 2	$SQE = \sum_{i=1}^N (y_i - \hat{y}_i)^2$	$QME = SQE / (N - 2)$
Total	N-1	$SQT = \sum_{i=1}^N y_i^2 - N\bar{Y}^2$	

$$F = \frac{SQR}{SQE/(N-2)}$$

$$R^2 = \frac{SQR}{SQT}$$

$\hat{\sigma}_\varepsilon^2$ Estimador da
variância do erro

Exemplo em Excel

Consumo de combustível de camiões de acordo com a carga transportada

Resultados gerados pelo Excel

	A	B	C	D
1	RESUMO DOS RESULTADOS			
2				
3	Estatística de regressão			
4	R múltiplo	0.87		
5	R-Quadrado	0.76		
6	R-quadrado ajustado	0.75		
7	Erro padrão	0.35		
8	Observações	40		
9				
10	ANOVA			
11		gl	SQ	MQ
12	Regressão	1	14.85	14.85
13	Resíduo	38	4.75	0.12
14	Total	39	19.60	
15				
16		Coefficientes	Erro padrão	Stat t
17	Interseção	8.8484	0.3840	23.0418
18	peso	-0.0604	0.0055	-10.9057
19				

$R^2 = 0,76$, ou seja, 76% da variação do rendimento é explicada pela equação de regressão $Y = 8,8484 - 0,0604X$

SQR

SQE

SQT

equação de regressão
 $Y = 8,8484 - 0,0604X$

Inferência estatística dos parâmetros da regressão

Inferência estatística

Modelo de regressão linear simples: $Y = \beta_0 + \beta_1 X + \varepsilon$

Teste t

Avalia a significância do coeficiente de regressão linear associado com uma determinada variável explicativa.

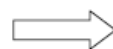
$H_0 : \beta_1 = 0$ (ausência do efeito)

$H_1 : \beta_1 \neq 0$ (presença do efeito)

Sob H_0

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{N-2}$$

Estatística teste



$t > t_{\text{crítico}}$ rejeita H_0

$t < t_{\text{crítico}}$ aceita H_0

$t_{\text{crítico}}$ é um valor tabelado para um nível de significância α , no Excel use `T.INV.2T(alfa;N-2)`

Inferência estatística

Exemplo da transportadora

Resultados gerados pelo Excel

A		Estatística teste				
$H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$		$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \sim t_{N-2}$				
5	R-Quadrado					
6	R-quadrado ajustado					
7	Erro padrão					
8	Observações	40				
9						
10	ANOVA					
11		gl	SQ	MQ	F	F de significação
12	Regressão	1	14.85	14.85	118.93	2.91E-13
13	Resíduo	38	4.75	0.12		
14	Total	39	19.60			
15						
16		Coeficientes	Erro padrão	Stat t	valor-P	95% inferiores
17	Interseção	8.8484	0.3840	23.0418	6.21E-24	8.0710
18	peso	-0.0604	0.0055	-10.9052	2.91E-13	-0.0716
19						

Ao nível de significância α de 5% o valor tabelado ($t_{\text{crítico}}$) de uma t com $(40-2) = 38$ graus de liberdade é 2,024 =INVT(0,05;38)

Valor absoluto do t calculado maior que $t_{\text{crítico}}$, logo H_0 é rejeitada.

$\hat{\beta}_1$

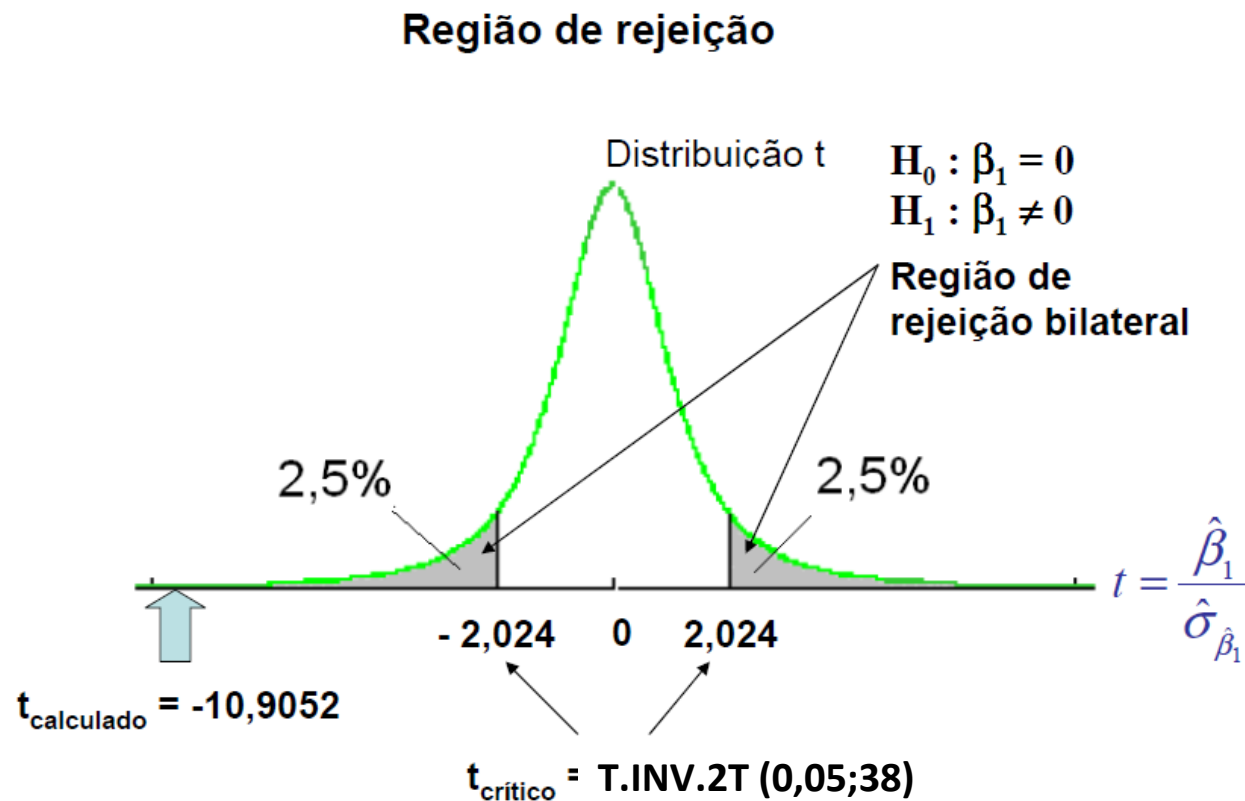
$\hat{\sigma}_{\hat{\beta}_1}$

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-0,0604}{0,0055} = -10,9052$$

t calculado

Inferência estatística

Aceitação ou rejeição da hipótese nula

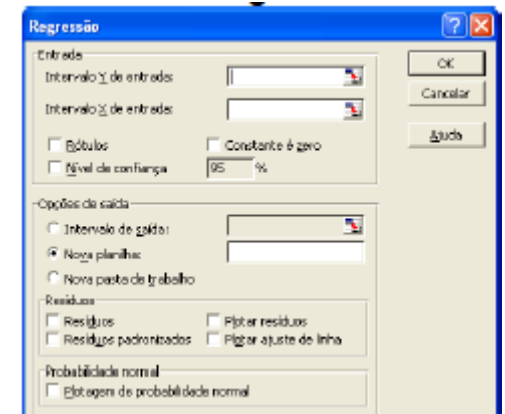
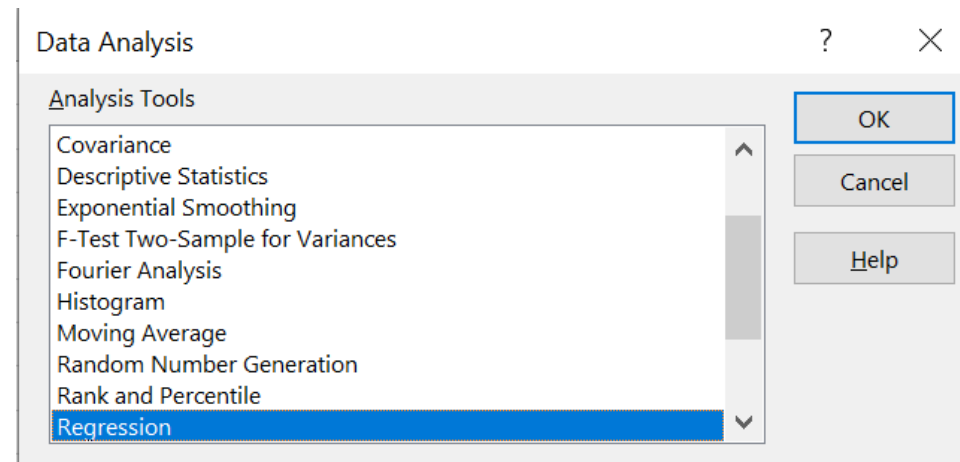


Exemplo de Regressão Linear Simples em Excel

1) Dados: Y (dependente); X (independente)

	E18		f_x
	A	B	
1	X	Y	
2	30	40	
3	20	34	
4	35	52	
5	40	49	
6	38	47	
7	18	21	
8	10	20	
9	15	27	
10	35	41	
11	24	48	
12			

2) No menu Data Analysis escolha Regression e surge a caixa de diálogo



Exemplo de Regressão Linear Simples em Excel

Intervalo com os valores da variável independente

Intervalo com os valores da variável dependente

Rótulos: nomes das variáveis

	A	B
1	X	Y
2	30	40
3	20	34
4	35	52
5	40	49
6	38	47
7	18	21
8	10	20
9	15	27
10	35	41
11	24	48

Marque se tem rótulo

Caixa de diálogo regressão

Regressão

Entrada

Intervalo Y de entrada: \$B\$1:\$B\$11

Intervalo X de entrada: \$A\$1:\$A\$11

☒ Rótulos

☐ Constante é zero

☐ Nível de confiança 95 %

Opções de saída

☐ Intervalo de saída:

☒ Nova planilha:

☐ Nova pasta de trabalho

Resíduos

☒ Resíduos

☐ Resíduos padronizados

☒ Plotar resíduos

☒ Plotar ajuste de linha

Probabilidade normal

☒ Plotagem de probabilidade normal

Gráfico dos resíduos contra a variável explicativa

Gráfico com os valores observados e previstos

Gráfico para avaliar se a hipótese de normalidade do erro é satisfeita

Apresenta a série de resíduos $Y - \hat{Y}$

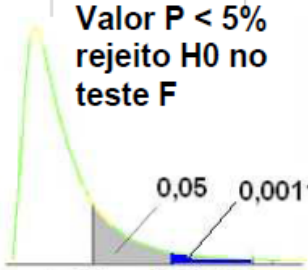
Grava resultados da regressão em uma nova planilha

Exemplo de Regressão Linear Simples em Excel

Resultados

RESUMO DOS RESULTADOS								
Estatística de regressão								
R múltiplo	0,8676	$\sqrt{R^2}$						
R-Quadrado	0,7527	R^2						
R-quadrado ajustado	0,7218							
Erro padrão	6,2432							
Observações	10							
ANOVA								
	gl	SQ	MQ	F	F de significação			
Regressão	1	949,0780	949,0780	24,3492	0,0011			
Resíduo	8	311,8220	38,9777					
Total	9	1260,9000						
	Coefficientes	Erro padrão	Stat t	valor-P	95% inferiores	95% superiores	Inferior 95,0%	Superior 95,0%
Interseção	12,0382	5,6005	2,1495	0,0638	-0,8766	24,9531	-0,8766	24,9531
X	0,9759	0,1978	4,9345	0,0011	0,5198	1,4320	0,5198	1,4320
RESULTADOS DE RESÍDUOS			RESULTADOS DE PROBABILIDADE					
Observação	Previsto(a) Y	Resíduos	Percentil	Y				
1	41,31570497	-1,315704967	5	20				
2	31,55654792	2,443452082	15	21				
3	46,19528349	5,804716508	25	27				
4	51,07486202	-2,074862017	35	34				
5	49,12303061	-2,123030607	45	40				
6	29,60471651	-8,604716508	55	41				
7	21,79739087	-1,797390868	65	47				
8	26,67696939	0,323030607	75	48				
9	46,19528349	-5,195283492	85	49				
10	35,46021074	12,53978926	95	52				

Valor P < 5%
rejeito H0 no
teste F



0,05 0,0011

5,3177 24,3492

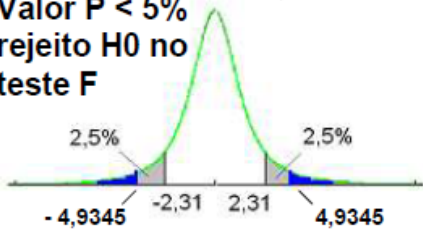
Valor P
 $P(F > 24,3492) = 0,0011$

Valor P
 $P(|t| > 4,9345) = 0,0011$

Valor P
 $P(|t| > 2,1495) = 0,0638$

Intervalo de
confiança

Valor P < 5%
rejeito H0 no
teste F



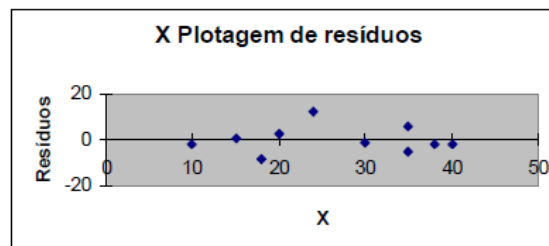
2,5% 2,5%

-4,9345 -2,31 2,31 4,9345

Valores para
a plotagem de
probabilidade
normal

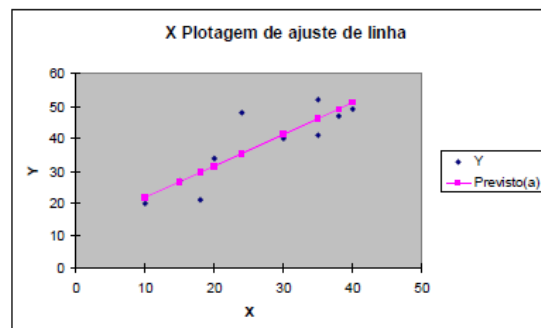
Exemplo de Regressão Linear Simples em Excel

Resultados

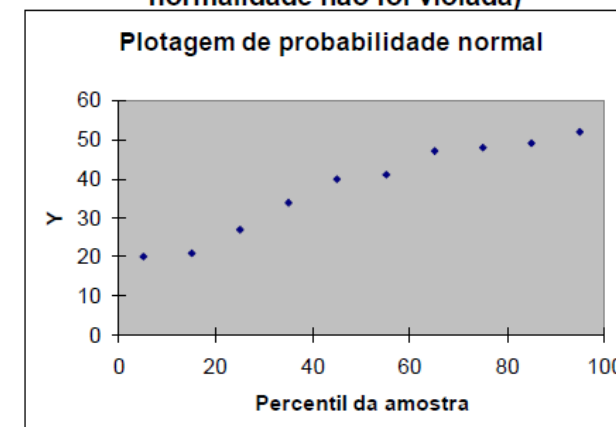


Útil na verificação da hipótese de
variância constante do erro

Valores observados contra valores estimados
Útil na avaliação da qualidade do ajuste



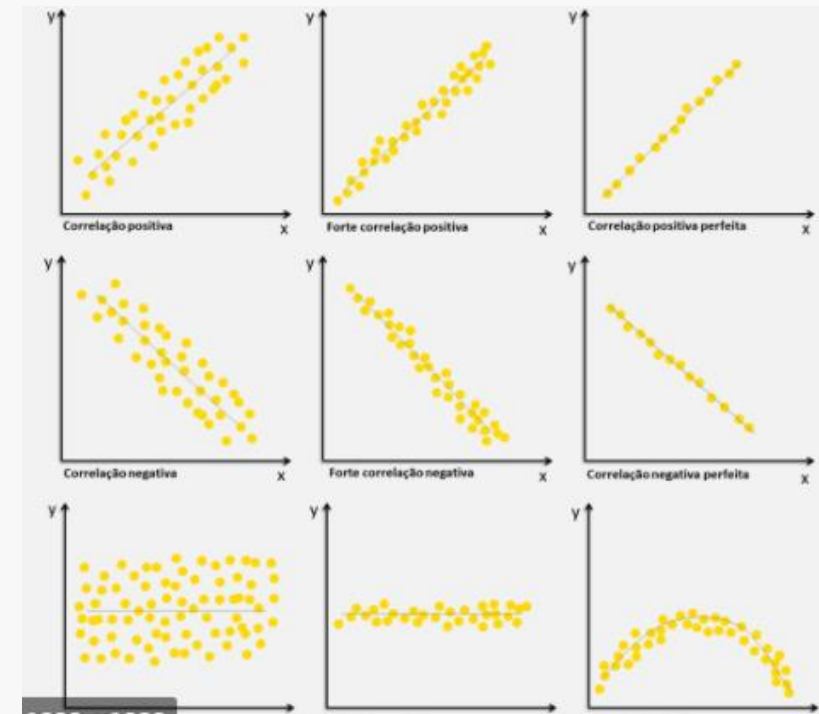
Útil na verificação da hipótese de
normalidade do erro (valores ao
redor de uma reta imaginária
indicam que a hipótese de
normalidade não foi violada)





02

Regressão Linear múltipla




Regressão Múltipla

A variável dependente é uma função linear de K variáveis independentes ($K \geq 2$)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{Ki} + \varepsilon_i \quad i=1, N$$

$\beta_1, \beta_2, \beta_3, \dots, \beta_k, \sigma^2$ são parâmetros do modelo que devem ser estimados


$$Y_i = [1 \quad X_{i,1} \quad \dots \quad X_{i,K}] \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon_i \quad i=1, N$$

Notação matricial $Y = X\beta + \varepsilon$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & \dots & X_{k1} \\ 1 & X_{12} & & X_{k2} \\ \vdots & & & \\ 1 & X_{1N} & & X_{kN} \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Hipóteses assumidas no modelo múltiplo

- H1)** A relação entre as variáveis é linear $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ki} + \varepsilon_i \quad i=1, n.$
- H2)** A variável explicativa X é fixa, ou seja, não é aleatória.
- H3)** As colunas da matriz X são linearmente independentes, ou seja, não há uma relação linear perfeita entre duas ou mais as variáveis explicativas.
- H4)** Erros tem média nula: $E(\varepsilon_i) = 0$ para todo $i=1, n.$
- H5)** Variância do erro é constante (homocedasticidade):
 $V(\varepsilon_i) = \sigma^2$ para todo $i=1, n.$
- H6)** Erros não correlacionados: $Cov(\varepsilon_i, \varepsilon_k) = 0$ para todo $i \neq k.$
- H7)** Erros tem distribuição Normal: $\varepsilon_i \sim N(0, \sigma^2)$ para todo $i=1, n.$

H2, H3, H4 e H5 - ε_i são independentes e identicamente distribuídos $N(0, \sigma^2)$

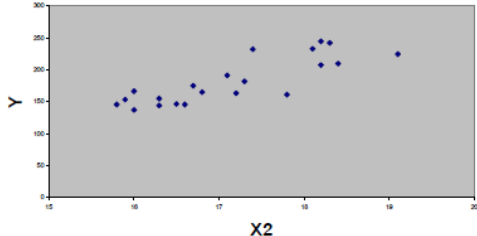
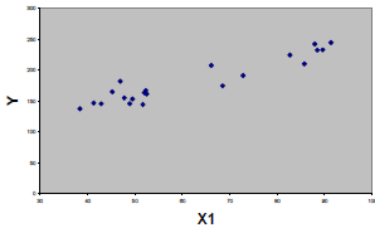
Exemplo

Uma empresa de artigos infantis opera em 21 cidades de médio porte. A empresa está analisando a possibilidade de expansão para outras cidades de médio porte e para isso deseja investigar se a vendas (Y) em uma localidade podem ser previstas com base no número de pessoas com até 16 anos de idades (X₁) e a renda per capita na localidade (X₂).

Valores expressos em milhares.

Atualmente a empresa está presente em 21 localidades (N = 21), cujos dados são apresentados na tabela ao lado:

X1	X2	Y
68,5	16,7	174,4
45,2	16,8	164,4
91,3	18,2	244,2
47,8	16,3	154,6
46,9	17,3	181,6
66,1	18,2	207,5
49,5	15,9	152,8
52	17,2	163,2
48,9	16,6	145,4
38,4	16	137,2
87,9	18,3	241,9
72,8	17,1	191,1
88,4	17,4	232
42,9	15,8	145,3
52,5	17,8	161,1
85,7	18,4	209,7
41,3	16,5	146,4
51,7	16,3	144
89,6	18,1	232,6
82,7	19,1	224,1
52,3	16	166,5

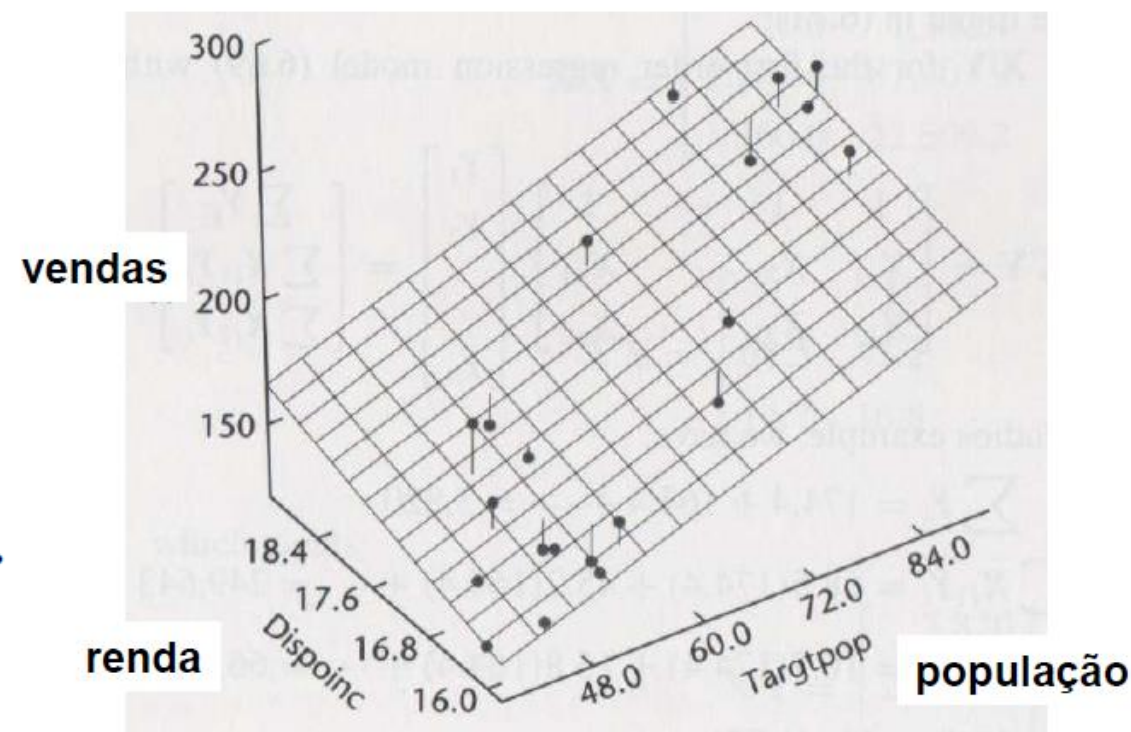


Equação que define o Plano

$$E(Y_i|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Equação estimada

$$Y = -68,86 + 1,45X_1 + 9,37X_2 + \varepsilon$$



ANOVA

Análise teórica

Causas de variação	Graus de liberdade	Soma dos quadrados	Quadrados médios
Regressão	K	$SQR = \hat{\beta}^T X^T Y - \left(\sum_{i=1}^N y_i \right)^2 / N$	$QMR = SQR / K$
Resíduos	N - (K+1)	$SQE = Y^T Y - \hat{\beta}^T X^T Y$	$QME = SQE / [N - (K + 1)]$
Total	N-1	$SQT = Y^T Y - \left(\sum_{i=1}^N y_i \right)^2 / N$	

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}$$

$$F = \frac{QMR}{QME} = \frac{SQR/k}{SQE/[N - (k + 1)]}$$

$$\overline{R}^2 = 1 - \left(1 - R^2 \right) \frac{N - 1}{N - k}$$

Exemplo da marca de artigos infantís

X_1	X_2	Y	\hat{Y}	$Y - \hat{Y}$	$Y - \bar{Y}$	$\hat{Y} - \bar{Y}$	$(Y - \hat{Y})^2$	$(\hat{Y} - \bar{Y})^2$	$(Y - \bar{Y})^2$
68,5	16,7	174,4	187,18	-12,78	-7,50	5,28	163,43	27,87	56,32
45,2	16,8	164,4	154,23	10,17	-17,50	-27,68	103,44	765,92	306,42
91,3	18,2	244,2	234,40	9,80	62,30	52,49	96,11	2755,36	3880,70
47,8	16,3	154,6	153,33	1,27	-27,30	-28,58	1,62	816,60	745,55
46,9	17,3	181,6	161,38	20,22	-0,30	-20,52	408,65	421,06	0,09
66,1	18,2	207,5	197,74	9,76	25,60	15,84	95,23	250,80	655,12
49,5	15,9	152,8	152,06	0,74	-29,10	-29,85	0,55	891,00	847,09
52	17,2	163,2	167,87	-4,67	-18,70	-14,04	21,78	197,07	349,87
48,9	16,6	145,4	157,74	-12,34	-36,50	-24,17	152,23	584,02	1332,60
38,4	16	137,2	136,85	0,35	-44,70	-45,06	0,13	2030,29	1998,52
87,9	18,3	241,9	230,39	11,51	60,00	48,48	132,54	2350,56	3599,43
72,8	17,1	191,1	197,18	-6,08	9,20	15,28	37,03	233,48	84,55
88,4	17,4	232	222,69	9,31	50,10	40,78	86,76	1663,08	2509,53
42,9	15,8	145,3	141,52	3,78	-36,60	-40,39	14,30	1631,06	1339,91
52,5	17,8	161,1	174,21	-13,11	-20,80	-7,69	171,96	59,16	432,84
85,7	18,4	209,7	228,12	-18,42	27,80	46,22	339,44	2136,21	772,58
41,3	16,5	146,4	145,75	0,65	-35,50	-36,16	0,43	1307,38	1260,59
51,7	16,3	144	159,00	-15,00	-37,90	-22,90	225,04	524,57	1436,77
89,6	18,1	232,6	230,99	1,61	50,70	49,08	2,60	2409,07	2570,01
82,7	19,1	224,1	230,32	-6,22	42,20	48,41	38,64	2343,65	1780,44
52,3	16	166,5	157,06	9,44	-15,40	-24,84	89,03	617,04	237,31
Total							2180,93	24015,28	26196,21

$$\hat{Y}_i = -68,8571 + 1,4546X_{1i} + 9,3655X_{2i}$$

SQE

SQR

SQT

Desenvolvimento do Exemplo

ANOVA

Fonte de variação	Soma dos quadrados (A)	Graus de liberdade (B)	Quadrado médio (C=A/B)	F
Regressão	SQR 24015,28	2	12007,64	12007,64 / 121.1626 = 99,1035
Resíduo	SQE 2180,93	N-3=18	121,1626	
Total	SQT 26196,21	N-1=20		

2 variáveis explicativas

3 coeficientes estimados
Por isso N – 3

O quadrado médio do resíduo é uma estimativa da variância do erro

Coeficiente de determinação R^2

$\hat{\sigma}_\varepsilon^2$

$R^2 = \frac{SQR}{SQT} = \frac{24015.28}{26196,21} = 0,917$

Inferência Estatística

Teste t

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

$$t = \frac{b_j}{\hat{\sigma}_{\beta_j}} \sim t_{N-(k+1)} \quad |t| \geq |t_{tabelado}| \Rightarrow \text{rejeita } H_0$$

K – variáveis explicativas)

**Teste de cada parâmetro
explicativo**

Teste F

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_1 : \text{pelo menos um } \beta_j \neq 0$$

$$F = \frac{SQR/k}{SQE/N-(k+1)} \quad |F| \geq |F_{tabelado}| \Rightarrow \text{rejeita } H_0$$

**Teste conjunto de todos os
parâmetros**

Inferência dos parâmetros

Exemplo do exercício em curso

Parâmetro B1

Teste t: Testa a significância do coeficiente de regressão linear associado com uma determinada variável explicativa.

$H_0 : b_1 = 0$ (ausência do efeito)

$H_1 : b_1 \neq 0$ (presença do efeito)

1) Estatística teste

$$t = \frac{\hat{b}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

2) Distribuição da estatística testes sob H_0

$$\frac{\hat{b}_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{N-3}$$

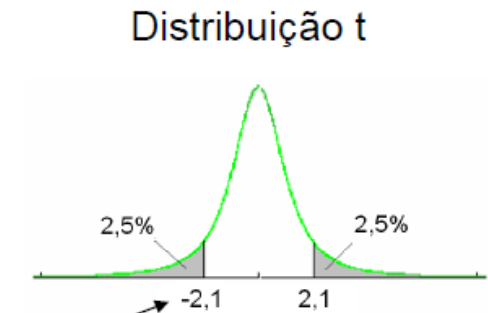
3) Valor da estatística teste na amostra observada ($t_{\text{calculado}}$)

$$t = \frac{1,4546}{0,2118} = 6,8682$$

4) t crítico ao nível de significância de 5% = 2,1 = TINV(0,05;18) no Excel

5) Conclusão

$t_{\text{calculado}} > t_{\text{crítico}}$ logo rejeita H_0



Inferência do Modelo

Parâmetro B2

Teste t: Testa a significância do coeficiente de regressão linear associado com uma determinada variável explicativa.

$H_0 : b_2 = 0$ (ausência do efeito)

$H_1 : b_2 \neq 0$ (presença do efeito)

1) Estatística teste

$$t = \frac{\hat{b}_2}{\hat{\sigma}_{\hat{\beta}_2}}$$

3) Valor da estatística teste na amostra observada ($t_{\text{calculado}}$)

$$t = \frac{9,3655}{4,0640} = 2,3045$$

2) Distribuição da estatística testes sob H_0

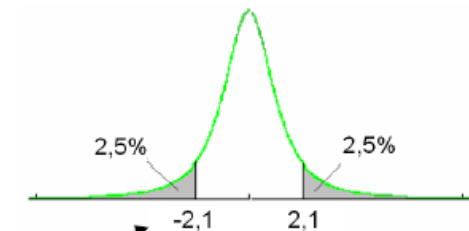
$$\frac{\hat{b}_2}{\hat{\sigma}_{\hat{\beta}_2}} \sim t_{N-3}$$

4) t crítico ao nível de significância de 5% = 2,1 =TINV(0,05;18) no Excel

5) Conclusão

$t_{\text{calculado}} > t_{\text{crítico}}$ logo rejeita H_0

Distribuição t



Inferência do Modelo na sua globalidade

Teste F

Teste F: Testa o efeito conjunto das variáveis explicativas sobre a variável dependente.

$H_0 : b_1 = b_2 = 0$ (não há regressão de Y em X_1 e X_2)

$H_1 : b_1 \neq 0$ ou $b_2 \neq 0$ (presença do efeito)

1) Estatística teste

$$F = \frac{\frac{SQR}{K}}{\frac{SQE}{N-(K+1)}}$$

3) Valor da estatística teste na amostra observada ($F_{\text{calculado}}$)

$$F = \frac{\frac{12.007,64}{2}}{\frac{121,1626}{21-(2+1)}} = 99,1035$$

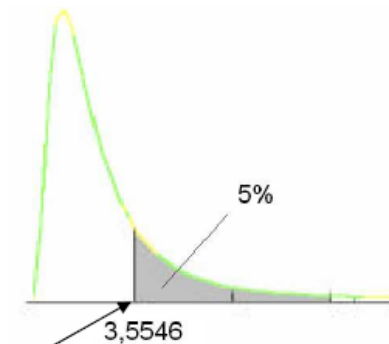
2) Distribuição da estatística testes sob H_0

$$\frac{\frac{SQR}{K}}{\frac{SQE}{N-(K+1)}} \sim F_{K,N-(K+1)}$$

4) F crítico ao nível de significância de 5% = 3,5546 =FINV(0,05;2;18) no Excel

5) Conclusão

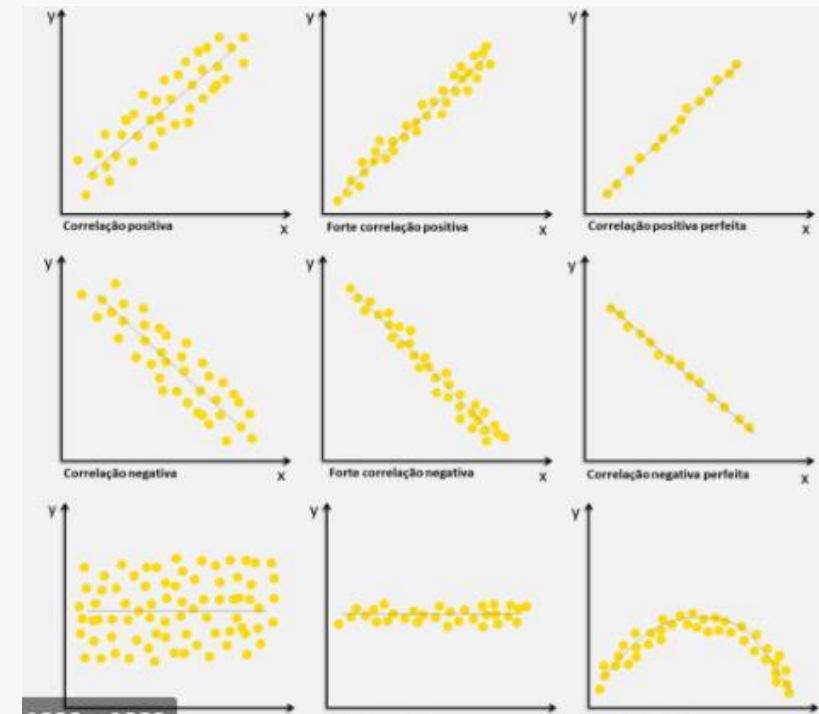
$F_{\text{calculado}} > F_{\text{crítico}}$ logo rejeita H_0





02

Regressão Logística



Regressão Logística

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, em função de uma ou mais variáveis independentes contínuas e/ou binárias

Suponha que queira-se analisar a ocorrência da apneia do sono, que é um distúrbio do sono potencialmente grave, em que a pessoa para de respirar, por alguns segundos, diversas vezes durante a noite. vamos considerar apenas dois: idade e peso.

A variável dependente é a ocorrência ou não da apneia do sono, ter apneia é igual a 1, não ter apneia é igual a 0. As variáveis independentes são a idade e o peso. Para este exemplo, o que a regressão logística propõe é que, a partir dessas informações, é possível gerar um modelo logístico que possa prever a probabilidade de uma pessoa ter apneia do sono, baseando-se no peso e idade desta pessoa.

Regressão Logística: Linearização com base na logaritmização

Exemplo

Se uma pessoa de 50 anos e 120 quilos tem probabilidade $p = 0,75$ de ter apneia. A probabilidade de não ter apneia é $1 - p$, logo, $1 - p = 0,25$. A probabilidade de p um evento ocorrer, contra ele não ocorrer, é uma razão de probabilidades, $0,75/0,25$ que é chamada de chance ($p/(1-p)$).

Assim temos que neste exemplo a proporção é de 3 para 1, isto significa que uma pessoa nessas características tem 3 vezes mais hipóteses de ter apneia do sono do que não ter.

Transformação linear:

P – sucesso

$1-p$ - insucesso

Se $P(Y=1) > 0,5$ então classifica-se $Y=1$

Se $P(Y=1) < 0,5$ então classifica-se $Y=0$

PG_MKT & Business Technologies

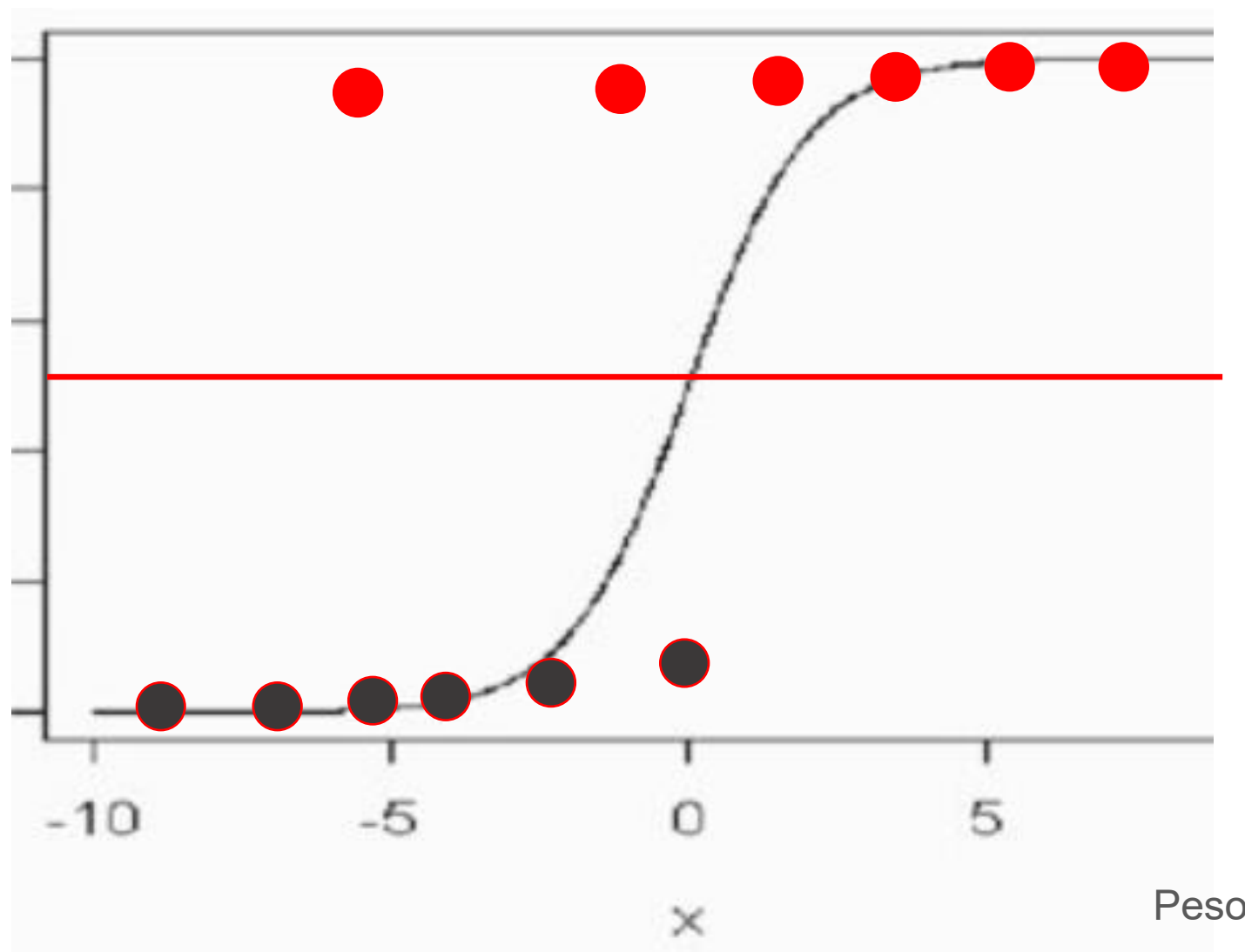
Models and Decision in Business Analytics

Regressão
Logística:
Obesidade e
Peso

Obeso

Obesidade

Não Obeso

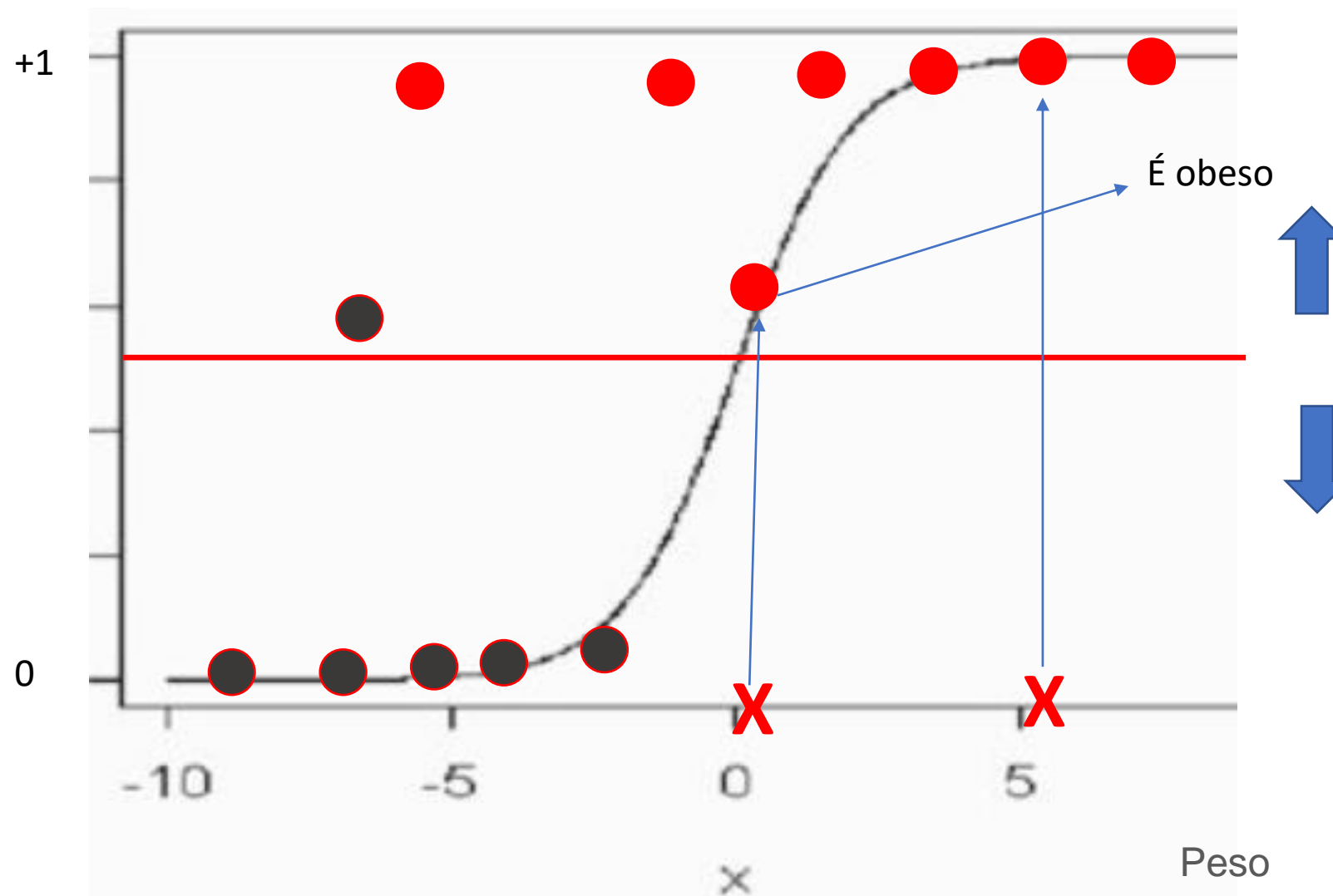


PG_MKT & Business Technologies

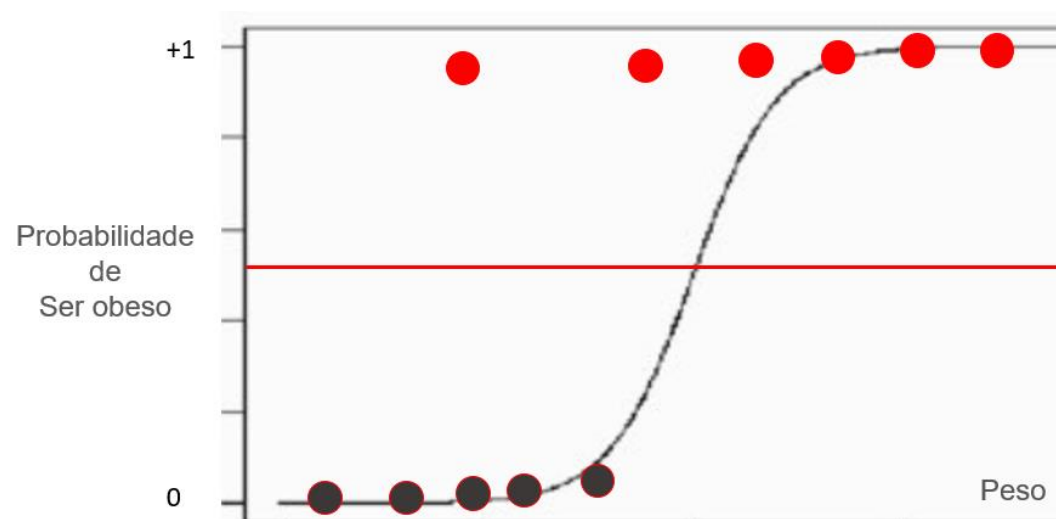
Models and Decision in Business Analytics

Regressão
Logística:
Obesidade e
Peso

Probabilidade
de
Ser obeso



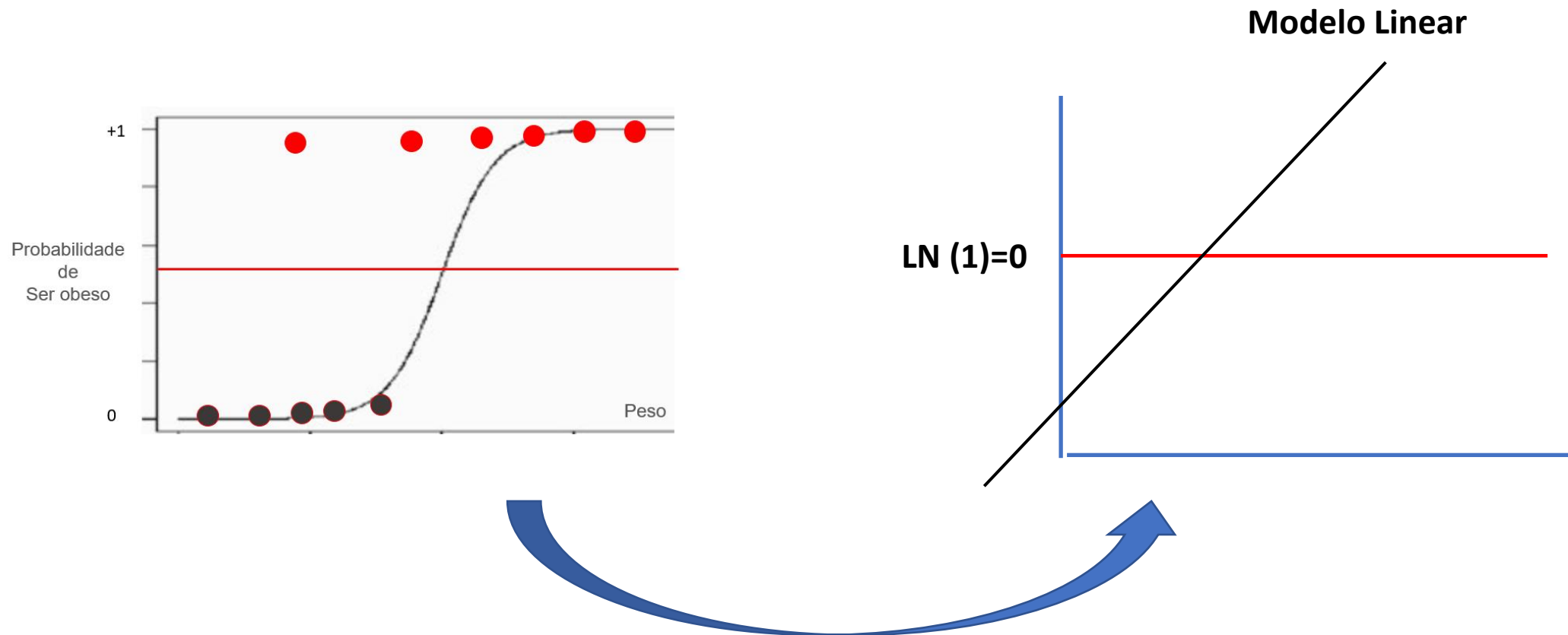
Regressão Logística: Obesidade e Peso



Linearização da regressão

$$\text{LN} (P / (1-P))$$

Regressão Logística: Obesidade e Peso



Regressão Logística: Linearização com base na logaritmização

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



$$p = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k)}}$$

Transformação linear:

P – sucesso

1-p - insucesso

Se $P(Y=1) > 0,5$ então classifica-se $Y=1$

Se $P(Y=1) < 0,5$ então classifica-se $Y=0$

Exemplo da Regressão Logística

Um grupo de investigadores estuda dados relativos a 875 partos. Óbito neonatal ou não, pré-termo (<37 semanas de gestação), mãe portadora ou não de diabetes.

O objetivo dos investigadores é verificar se a probabilidade de óbito neonatal é função das variáveis explanatórias.

Dados conhecidos no levantamento para o estudo

Número de partos	Número de óbitos neonatais	Percentual de óbitos neonatais	Pré-termo X1	Mãe com diabetes X2
10	1	10,00%	1	1
14	2	14,29%	0	1
181	16	8,84%	1	0
670	5	0,75%	0	0
1=sim 0=não				

Aplicação da Regressão Logística

Número de partos	Número de óbitos	p	1-p	W=p/(1-p)	Lógite (p) Ln (w)
10	1	0,100	0,900	0,111	-2,20
14	2	0,143	0,857	0,167	-1,79
181	16	0,0884	0,912	0,097	-2,33
670	5	0,00746	0,993	0,008	-4,89

$$\text{LN } (P/(1-P)) = - 4,15 + 1,076 X_1 + 1,617 X_2$$

$$P = \frac{1}{1 + e^{(- 4,15 + 1,076 X_1 + 1,617 X_2)}}$$



02

Árvores de Decisão



Conceito de Árvore de Decisão (entropia e ganho de informação)

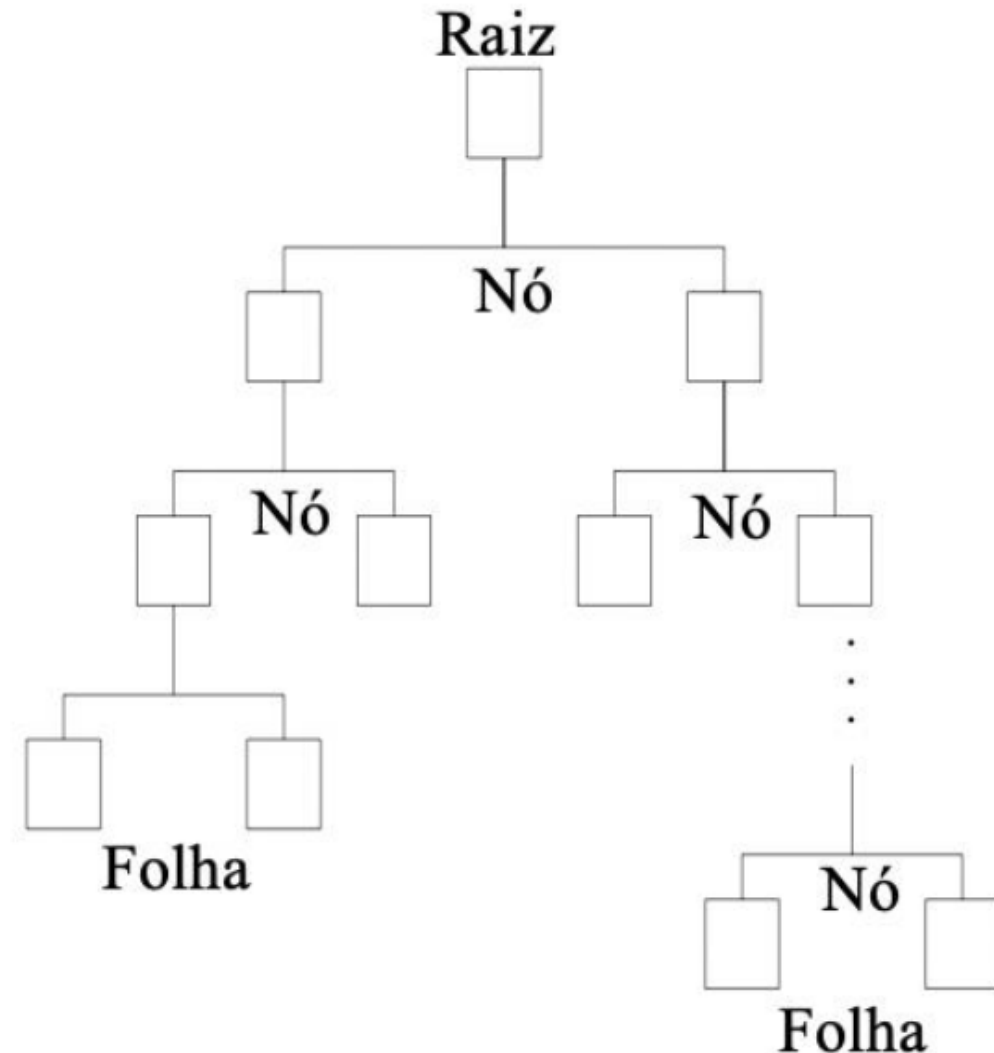
Uma árvore de decisão é uma ferramenta de suporte à tomada de decisão que usa um gráfico no formato de árvore e demonstra visualmente as condições e as probabilidades para se chegar a resultados.

Estrutura de uma Árvore de Decisão

Exemplo de árvore

O método consiste na continua subdivisão de um espaço amostral em classe menores por meio de testes

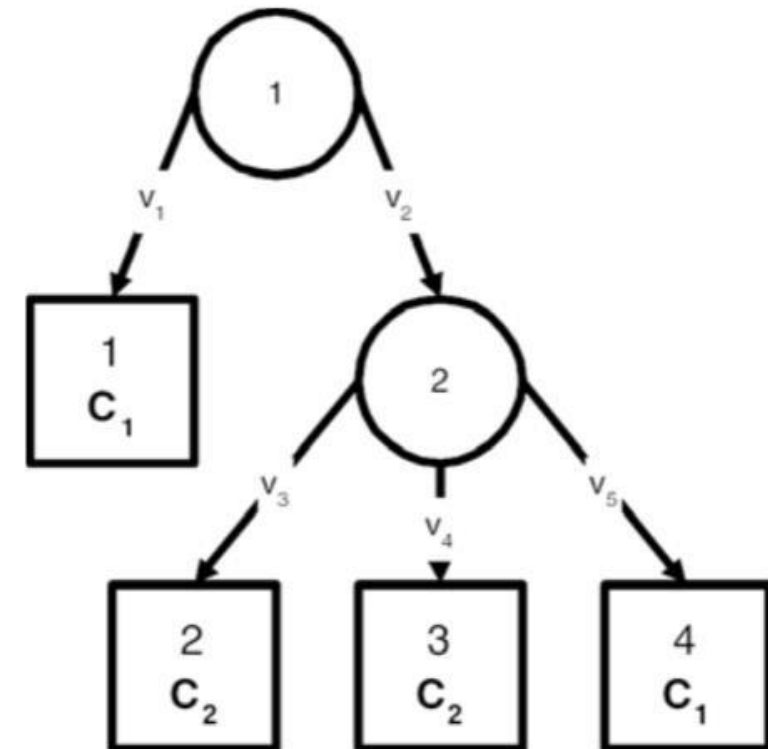
- Algoritmo para tomar decisões (ou classificar)
- Modo de representar conhecimento



Desenvolvimento

A figura ilustra um árvore de decisão não binária, isto é, o nó 2 possui três ramos.

Neste tipo de árvore, um teste realizado em um nó resulta na divisão de dois ou mais conjuntos disjuntos que cobrem todas as possibilidades, isto é, todo novo caso deve pertencer a um dos subconjuntos disjuntos.



O processo de aprendizagem da estrutura de uma árvore de decisão é conhecido com indução ou regras. A indução busca padrões em informações disponíveis com o propósito de inferir conclusões racionais.

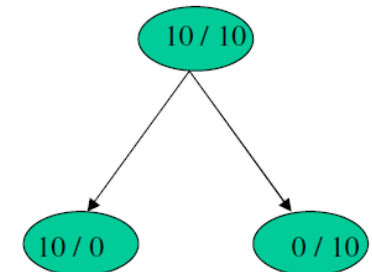
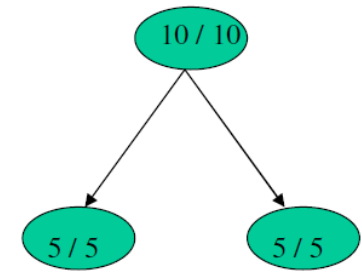


Como medir a *habilidade* de um dado atributo discriminar as classes?

- Existem muitas medidas.

Todas concordam em dois pontos:

- Uma divisão que mantém as proporções de classes em todas as partições é inútil.
- Uma divisão onde em cada partição todos os exemplos são da mesma classe tem utilidade máxima



Entropia

Entropia é uma medida da aleatoriedade de uma variável (da sua impureza)

A entropia de uma variável nominal X que pode tomar i valores:

$$entropia(X) = - \sum_i p_i * \log_2 p_i$$

Entropia do Sistema

- A entropia tem máximo ($\log_2 i$) se $p_i = p_j$ para qualquer $i \neq j$
- A entropia(x) = 0 se existe um i tal que $p_i = 1$
- É assumido que $0 * \log_2 0 = 0$

Criação de um Nó

Para definir qual o melhor critério dentre todos os possíveis é feito o cálculo do ganho de informação, que consiste na análise da homogeneidade das subclasses criadas, escolhendo, assim o critério que traga um maior ganho de informação (*Ganho*):

$$Ganho(T) = Inf(T) - \sum_{t=1}^m \frac{|T_t|}{|T|} * Inf(T_t)$$

O ganho de informação mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo.


A construção de uma árvore de decisão é guiada pelo objectivo de diminuir a entropia ou seja a aleatoriedade - dificuldade de previsão da variável objectivo

Exercício: Construção da Árvore

Dados disponíveis de diferentes amostras

- 4 atributos:
Céu,
Temperatura
Humidade
Vento

Nº exemplar	Céu	Temperatura	Humidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
3	nublado	alta	alta	não	joga
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
7	nublado	baixa	normal	sim	joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
10	chuva	suave	normal	não	joga
11	sol	suave	normal	sim	joga
12	nublado	suave	alta	sim	joga
13	nublado	alta	normal	não	joga
14	chuva	suave	alta	sim	não joga



9 hipóteses de
haver jogo e 5
hipóteses de
não ocorrer

Exemplo: Construção da Árvore

O objetivo do cálculo da entropia está na classificação booleana (jogar golfe × não jogar golfe), em que há 14 exemplos, 9 positivos e 5 negativos, ou seja, $T = 9+, 5-$.

$$\begin{aligned}\text{info}(T) &= -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_n \log_2 p_n \\ &= -\left(\frac{9}{14}\right) \log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \left(\frac{5}{14}\right) \\ &= 0,940\end{aligned}$$

Após calcular a entropia do sistema, busca-se qual atributo possui melhor ganho de informação.

Exemplo: Construção da Árvore

Ganho de informação (CÉU)

O atributo céu pode assumir 3 valores (sol, nublado e chuva).

$$T_{\text{sol}} = [2+, 3-], T_{\text{nublado}} = [4+, 0-] \text{ e } T_{\text{chuva}} = [3+, 2-]$$

$$\text{info}(\text{sol}) = -\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) = 0,97094$$

$$\text{info}(\text{nublado}) = -\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) = 0 \quad \text{info}(\text{chuva}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)$$

$$\begin{aligned} \text{Logo, } \text{Ganho}(\text{info}(T), \text{céu}) &= 0,940 - \left(\frac{5}{14}\right) \cdot \text{info}(\text{sol}) - \left(\frac{4}{14}\right) \cdot \text{info}(\text{nublado}) - \left(\frac{5}{14}\right) \cdot \text{info}(\text{chuva}) \\ &= 0,940 - \left(\frac{5}{14}\right) \cdot 0,97094 - \left(\frac{4}{14}\right) \cdot 0 - \left(\frac{5}{14}\right) \cdot 0,97094 = 0,2464 \end{aligned}$$

Exemplo: Construção da Árvore

$$T_{\text{alta}} = [3+, 2-], T_{\text{suave}} = [3+, 1-] \text{ e } T_{\text{baixa}} = [3+, 2-]$$

Ganho de informação (TEMPERATURA)

O atributo temperatura pode assumir 3 valores (alta, suave e baixa).

$$\text{info}(\text{alta}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0,97094$$

$$\text{info}(\text{baixa}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0,97094$$

$$\text{info}(\text{suave}) = -\left(\frac{3}{4}\right)\log_2\left(\frac{3}{4}\right) - \left(\frac{1}{4}\right)\log_2\left(\frac{1}{4}\right) = 0,811$$

Ganho Informação

$$\begin{aligned} \text{Ganho}(\text{info}(T), \text{temperatura}) &= 0,940 - \left(\frac{5}{14}\right) \cdot \text{info}(\text{alta}) - \left(\frac{4}{14}\right) \cdot \text{info}(\text{suave}) - \left(\frac{5}{14}\right) \cdot \text{info}(\text{baixa}) \\ &= 0,940 - \left(\frac{5}{14}\right) \cdot 0,97094 - \left(\frac{4}{14}\right) \cdot 0,811 - \left(\frac{5}{14}\right) \cdot 0,97094 = 0,015 \end{aligned}$$

Exemplo: Construção da Árvore

$$T_{\text{alta}} = [3+, 4-] \text{ e } T_{\text{baixa}} = [6+, 1-]$$

Ganho de informação – (HUMIDADE)

O atributo umidade pode assumir 2 valores (alta e baixa).

$$\text{info}(\text{alta}) = -\left(\frac{3}{7}\right)\log_2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right)\log_2\left(\frac{4}{7}\right) = 0,985228$$

$$\text{info}(\text{baixa}) = -\left(\frac{6}{7}\right)\log_2\left(\frac{6}{7}\right) - \left(\frac{1}{7}\right)\log_2\left(\frac{1}{7}\right) = 0,591672$$

**Ganho
Informação**

$$\begin{aligned} \text{Ganho}(\text{info}(T), \text{umidade}) &= 0,940 - \left(\frac{7}{14}\right) \cdot \text{info}(\text{alta}) - \left(\frac{7}{14}\right) \cdot \text{info}(\text{baixa}) \\ &= 0,940 - \left(\frac{7}{14}\right) \cdot 0,985228 - \left(\frac{7}{14}\right) \cdot 0,591672 = 0,151 \end{aligned}$$

Exemplo: Construção da Árvore

$$T_{\text{sim}} = [3+, 3-], T_{\text{não}} = [6+, 2-]$$

Ganho de informação – (VENTO)

O atributo vento pode assumir 2 valores (sim e não).

$$\text{info}(\text{sim}) = -\left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right) = 1$$

$$\text{info}(\text{não}) = -\left(\frac{6}{8}\right)\log_2\left(\frac{6}{8}\right) - \left(\frac{2}{8}\right)\log_2\left(\frac{2}{8}\right) = 0,811278$$

**Ganho
Informação**

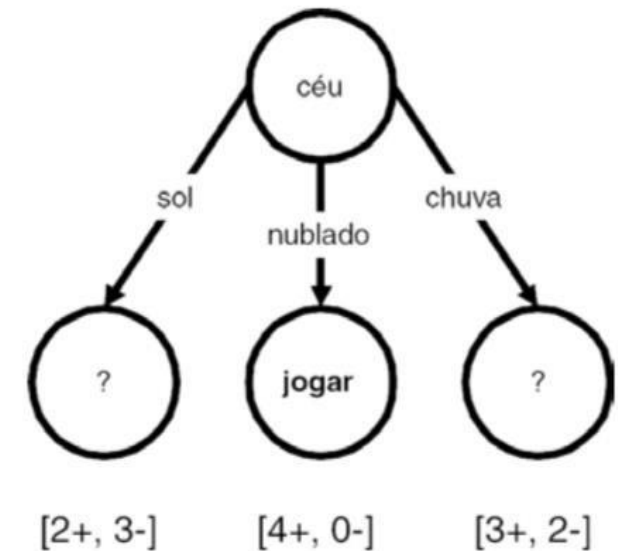
$$\begin{aligned} \text{Ganho}(\text{info}(T), \text{vento}) &= 0,940 - \left(\frac{8}{14}\right) \cdot \text{info}(\text{sim}) - \left(\frac{6}{14}\right) \cdot \text{info}(\text{não}) \\ &= 0,940 - \left(\frac{6}{14}\right) \cdot 1 - \left(\frac{8}{14}\right) \cdot 0,811278 = 0,047841 \end{aligned}$$

Exemplo: Construção da Árvore

Escolhe-se o atributo (céu) de maior ganho de informação para ser o nó raiz da árvore.

<i>Ganho info T , céu</i>	= 0, 2464
<i>Ganho info T , temperatura</i>	= 0,015
<i>Ganho info T , umidade</i>	= 0,151
<i>Ganho info T , vento</i>	= 0,047841

Os ramos sol e chuva ainda estão indefinidos, e processo deve continuar no próximo nível da árvore.



Exemplo: Construção da Árvore

Céu = sol

$T1 = \{1, 2, 8, 9, 11\}$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

Céu = nublado

$T2 = \{3, 7, 12, 13\}$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
3	nublado	alta	alta	não	joga
7	nublado	baixa	normal	sim	joga
12	nublado	suave	alta	sim	joga
13	nublado	alta	normal	não	joga

Céu = chuva

$T3 = \{4, 5, 6, 10, 14\}$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
10	chuva	suave	normal	não	joga
14	chuva	suave	alta	sim	não joga

PG_MKT & Business Technologies

Models and Decision in Business Analytics



Exemplo: Construção da Árvore

Céu = sol (processo de indução para este ramo da árvore) Ganho de informação – TEMPERATURA

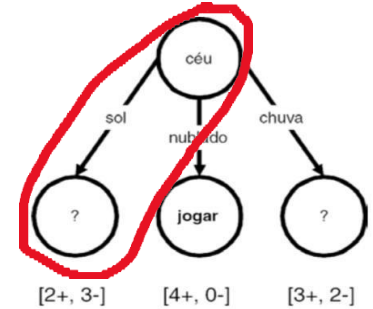
$$T_{alta} = [2+, 0-], T_{suave} = [1+, 1-] \text{ e } T_{baixa} = [0+, 1-]$$

$$\text{info}(alta) = -\left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) = 0$$

$$\text{info}(suave) = -\left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) - \left(\frac{1}{2}\right) \log_2 \left(\frac{1}{2}\right) = 1$$

$$\text{info}(baixa) = -\left(\frac{1}{1}\right) \log_2 \left(\frac{1}{1}\right) - \left(\frac{1}{1}\right) \log_2 \left(\frac{1}{1}\right) = 0$$

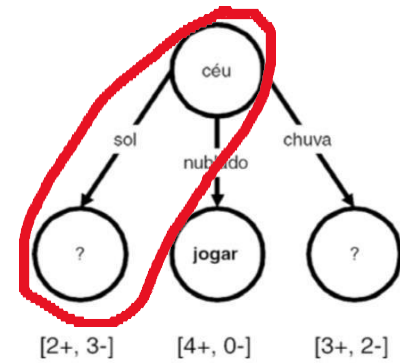
$$\begin{aligned} \text{Logo, } \text{Ganho}(\text{info}(\text{sol}), \text{temperatura}) &= 0,97094 - \left(\frac{2}{5}\right) \cdot \text{info}(alta) - \left(\frac{2}{5}\right) \cdot \text{info}(suave) - \left(\frac{1}{5}\right) \cdot \text{info}(baixa) \\ &= 0,97094 - \left(\frac{2}{5}\right) \cdot 0 - \left(\frac{2}{5}\right) \cdot 1 - \left(\frac{1}{5}\right) \cdot 0 \\ &= 0.57094 \end{aligned}$$



Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

PG_MKT & Business Technologies

Models and Decision in Business Analytics



Exemplo: Construção da Árvore

Céu = sol (processo de indução para este ramo da árvore) Ganho de informação – HUMIDADE

$$T_{alta} = [3+, 0-] \text{ e } T_{baixa} = [0+, 2-]$$

$$\text{info}(alta) = -\left(\frac{3}{3}\right) \log_2\left(\frac{3}{3}\right) = 0$$

$$\text{info}(baixa) = -\left(\frac{2}{2}\right) \log_2\left(\frac{2}{2}\right) = 0$$

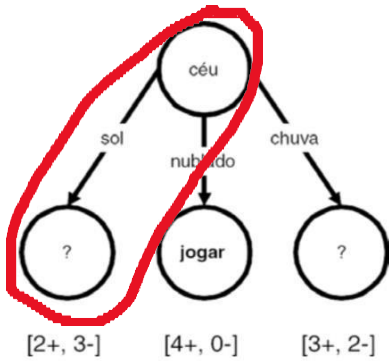
Logo,

$$\begin{aligned} \text{Ganho}(\text{info}(s), \text{umidade}) &= 0,97094 - \left(\frac{3}{5}\right) \cdot \text{info}(alta) - \left(\frac{2}{5}\right) \cdot \text{info}(baixa) \\ &= 0,97094 - \left(\frac{3}{5}\right) \cdot 0 - \left(\frac{2}{5}\right) \cdot 0 \\ &= 0,97094 \end{aligned}$$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

PG_MKT & Business Technologies

Models and Decision in Business Analytics



Exemplo: Construção da Árvore

Céu = sol (processo de indução para este ramo da árvore) Ganho de informação – VENTO

$$T_{\text{sim}} = [1+, 1-], T_{\text{não}} = [2+, 1-]$$

$$\text{info}(\text{sim}) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$\text{info}(\text{não}) = -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) = 0,918295$$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

Logo,

Logo,

$$\begin{aligned} \text{Ganho}(\text{info}(\text{sol}), \text{vento}) &= 0,97094 - \left(\frac{2}{5}\right) \cdot \text{info}(\text{sim}) - \left(\frac{3}{5}\right) \cdot \text{info}(\text{não}) \\ &= 0,97094 - \left(\frac{2}{5}\right) \cdot 1 - \left(\frac{3}{5}\right) \cdot 0,918295 \\ &= 0.019963 \end{aligned}$$

Exemplo: Construção da Árvore

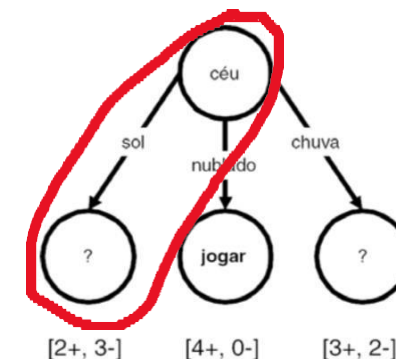
Céu = sol (processo de indução para este ramo da árvore)

Ganho info sol , temperatura = 0,57094

***Ganho info sol , humidade* = 0,97094**

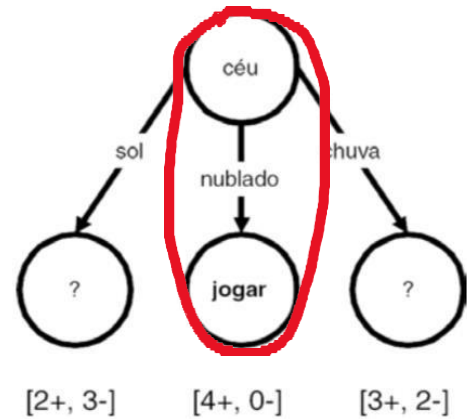
Ganho info sol , vento = 0,019963

Examinando os ganhos verifica-se que o atributo com maior ganho de informação é a umidade, o qual deve ser o nó seguinte da árvore neste ramo.



Exemplo: Construção da Árvore

Céu = nublado (processo de indução para este ramo da árvore)

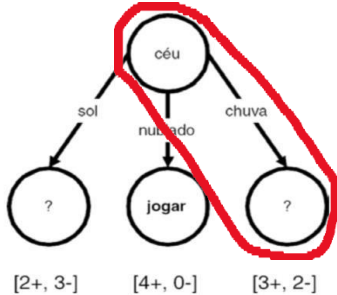


Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
3	nublado	alta	alta	não	joga
7	nublado	baixa	normal	sim	joga
12	nublado	suave	alta	sim	joga
13	nublado	alta	normal	não	joga

Observa-se que todas as amostras contidas nesse subconjunto pertencem somente a uma classe (jogar). Neste caso, o processo de indução acaba para este subconjunto e um nó folha é gerado.

Exemplo: Construção da Árvore

Céu = chuva (proc. de indução para este ramo da árvore) Ganho de informação – TEMPERATURA



$$T_{alta} = [2+, 0-], T_{suave} = [1+, 1-] \text{ e } T_{baixa} = [0+, 1-]$$

$$\text{info}(alta) = -\left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) = 0$$

$$\text{info}(suave) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$\text{info}(baixa) = -\left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) - \left(\frac{1}{1}\right)\log_2\left(\frac{1}{1}\right) = 0$$

Logo,

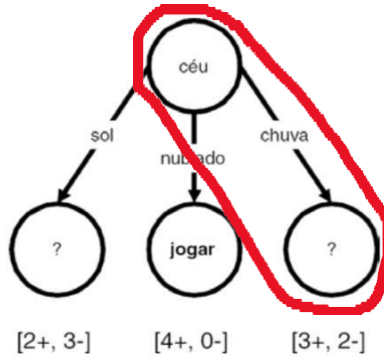
$$\begin{aligned} \text{Ganho}(\text{info}(chuva), temperatura) &= 0,97094 - \left(\frac{1}{5}\right) \cdot \text{info}(alta) - \left(\frac{2}{5}\right) \cdot \text{info}(suave) - \left(\frac{2}{5}\right) \cdot \text{info}(baixa) \\ &= 0,97094 - \left(\frac{1}{5}\right) \cdot 0 - \left(\frac{2}{5}\right) \cdot 1 - \left(\frac{2}{5}\right) \cdot 1 \end{aligned}$$

$$= 0,17090$$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
1	sol	alta	alta	não	não joga
2	sol	alta	alta	sim	não joga
8	sol	suave	alta	não	não joga
9	sol	baixa	normal	não	joga
11	sol	suave	normal	sim	joga

PG_MKT & Business Technologies

Models and Decision in Business Analytics



Exemplo: Construção da Árvore

Céu = chuva (proc. de indução para este ramo da árvore) Ganho de informação – HUMIDADE

$$T_{alta} = [1+, 1-], \quad T_{normal} = [2+, 1-]$$

$$\text{info}(alta) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

$$\text{info}(normal) = -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) = 0,9182958$$

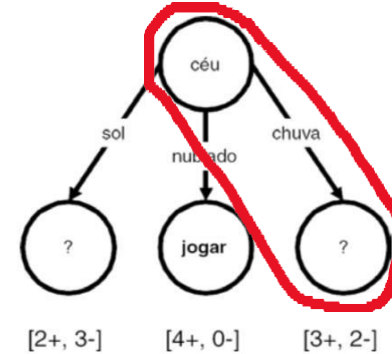
Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
10	chuva	suave	normal	não	joga
14	chuva	suave	alta	sim	não joga

**Ganho
Informação**

$$\begin{aligned} \text{Ganho}(\text{info}(chuva), \text{umidade}) &= 0,97094 - \left(\frac{2}{5}\right) \cdot 1 - \left(\frac{3}{5}\right) \cdot 0,9182958 \\ &= 0,019962 \end{aligned}$$

Exemplo: Construção da Árvore

Céu = chuva (proc. de indução para este ramo da árvore) Ganho de informação – **VENTO**



$$T_{\text{sim}} = [2+, 0-], T_{\text{não}} = [0+, 3-]$$

$$\text{info}(\text{sim}) = -\left(\frac{2}{2}\right) \log_2 \left(\frac{2}{2}\right) = 0$$

$$\text{info}(\text{não}) = -\left(\frac{3}{3}\right) \log_2 \left(\frac{3}{3}\right) = 0$$

Nº exemplar	Céu	Temperatura	Umidade	Vento	Classe
4	chuva	alta	alta	não	joga
5	chuva	baixa	normal	não	joga
6	chuva	baixa	normal	sim	não joga
10	chuva	suave	normal	não	joga
14	chuva	suave	alta	sim	não joga

$$\text{Logo, } \text{Ganho}(\text{info}(\text{chuva}), \text{vento}) = 0,97094 - \left(\frac{2}{5}\right) \cdot \text{info}(\text{sim}) - \left(\frac{3}{5}\right) \cdot \text{info}(\text{não})$$

$$= 0,97094 - \left(\frac{2}{5}\right) \cdot 0 - \left(\frac{3}{5}\right) \cdot 0$$

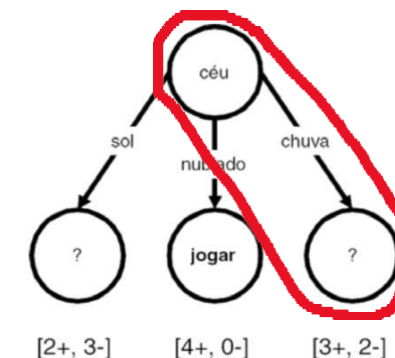
$$= 0.97094$$

Exemplo: Construção da Árvore

Céu = chuva (proc. de indução para este ramo da árvore)

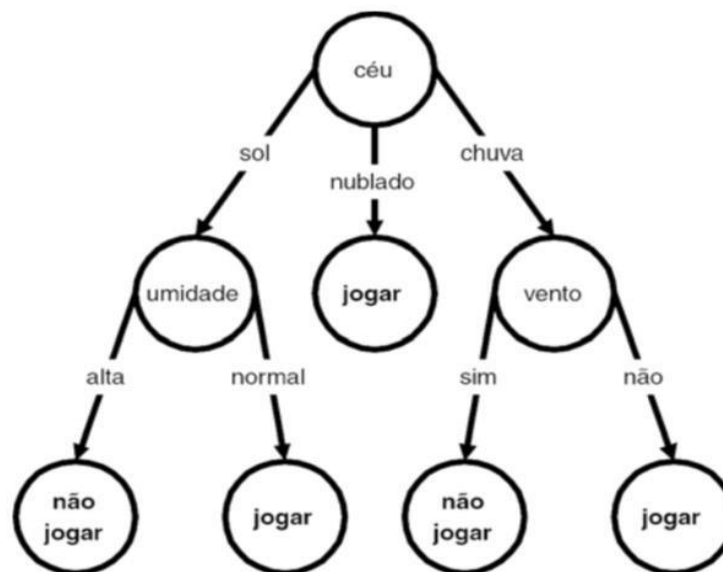
<i>Ganho info chuva , temperatura</i>	= 0,17090
<i>Ganho info chuva , Humidade</i>	= 0,019962
<i>Ganho info chuva , vento</i>	= 0, 97094

Examinando os ganhos verifica-se que o atributo com maior ganho de informação é o vento, o qual deve ser o nó seguinte na árvore.



Exemplo: Construção da Árvore

Árvore de decisão final para o conjunto de treinamento:



Observa-se que o atributo temperatura não foi selecionado para fazer parte da árvore (irrelevante para a tarefa de classificação, neste caso).

Obrigado!
(luisflcosta@sapo.pt)