



SHIFT

 FIAP





PYTHON JOURNEY

MACHINE & DEEP LEARNING

REGRESSÃO LINEAR



LIMITAÇÕES DOS TESTES DE HIPÓTESE

- Estabelecem se existe associação entre duas variáveis, mas...
- Não quantificam a força da associação; e
- Não permitem representar a relação existente sob uma forma funcional.



RELEMBRANDO: COEFICIENTE DE CORRELAÇÃO

Coeficiente de correlação linear de Pearson

Valor numérico que mede a intensidade da associação linear existente entre as duas variáveis a partir de uma série de observações.

O coeficiente de correlação assume valores entre -1 e 1.

Valores próximos de 1 indicam uma forte relação linear positiva.


Valores próximos de -1 indicam uma forte relação linear negativa.



COEFICIENTE DE CORRELAÇÃO

Correlação indica a força e a direção do relacionamento linear entre duas variáveis aleatórias. No uso estatístico geral, correlação se refere à medida da relação entre duas variáveis, embora correlação não implique causalidade. Neste sentido, existem vários coeficientes medindo o grau de correlação adaptados à natureza dos dados.

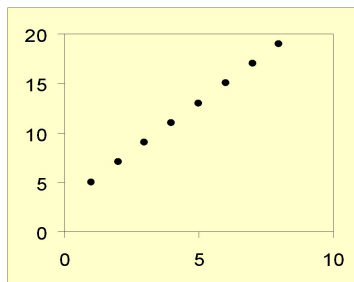
Covariância



$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

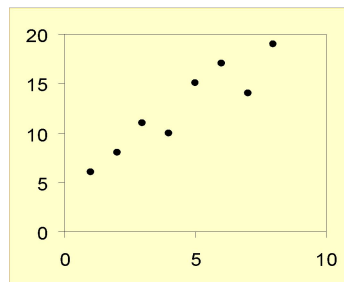


RELEMBRANDO: COEFICIENTE DE CORRELAÇÃO

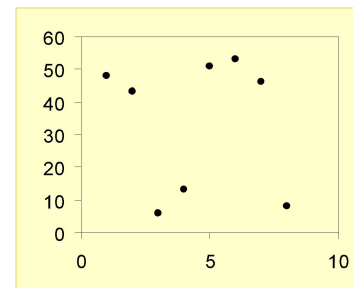


$r = +1$

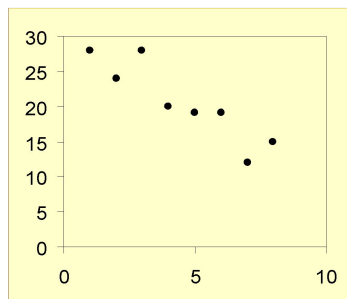
Relação
perfeita



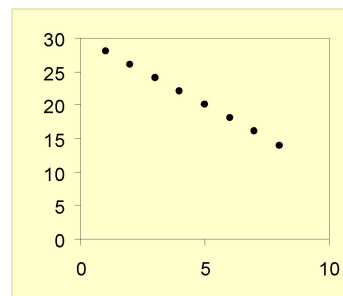
$r = +0,80$



$r = 0$



$r = -0,80$



$r = -1$

Relação
perfeita



REGRESSÃO LINEAR

SIMPLES / MÚLTIPLA

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

PREVISÕES sobre o comportamento futuro de um fenômeno atual extrapolam-se para o futuro o comportamento presente das variáveis:

Ex.: Prever a população de uma cidade no futuro.

Ex.: Prever a natalidade infantil para o ano 2050.

Ex.: Prever a demanda futura por habitação.



REGRESSÃO LINEAR SIMPLES / MÚLTIPLA

SIMULAR os efeitos de uma variável X sobre uma variável Y avalia-se as relações de causa-efeito entre 2 variáveis.

Ex.: Simular os efeitos sobre a segurança na cidade (Y) em função do aumento do policiamento ostensivo nas ruas (X) .

Ex.: Simular o efeito sobre o trânsito (Y) de uma cidade em função da elevação do preço da gasolina (X).



REGRESSÃO LINEAR

SIMPLES / MÚLTIPLA

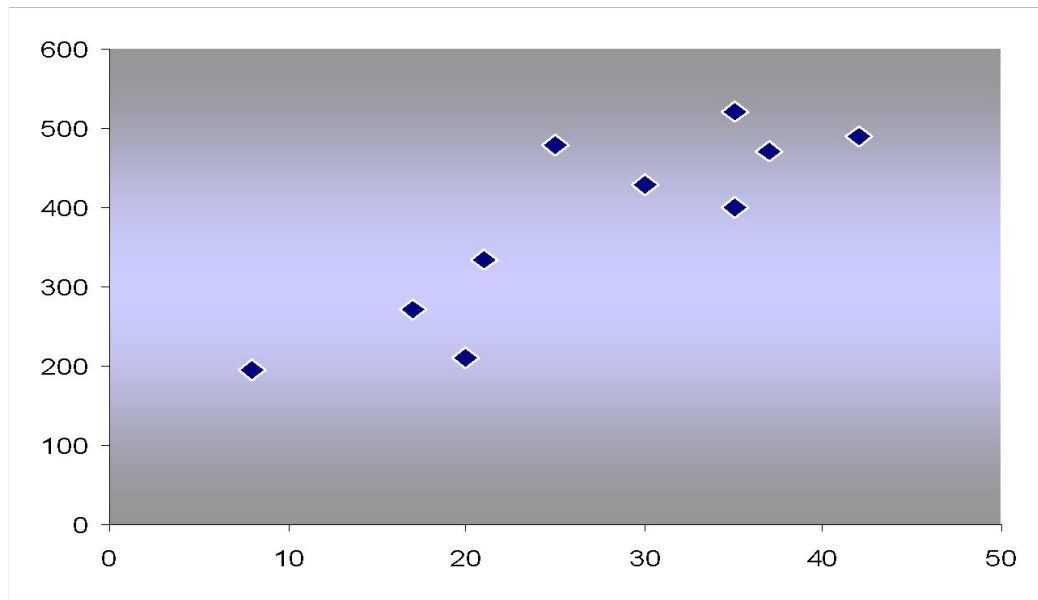
Observação	X	Y
1	30	4300
2	21	3350
3	35	5200
4	42	4900
5	37	4700
6	20	2100
7	8	1950
8	17	2700
9	35	4000
10	25	4800

Relação Direta

Idade X Renda mensal



REGRESSÃO LINEAR SIMPLES / MÚLTIPLA



Relação Direta

Idade X Renda mensal



MODELO LINEAR

A Análise de Regressão é o processo matemático para calcular os parâmetros “a” e “b” de uma função $f(X)$.

$$Y = a + b X$$

Estes parâmetros determinam as características da função que relaciona ‘Y’ com ‘X’.

No caso do modelo linear, esta função é representada pela chamada reta de regressão.

A regressão significa que os pontos plotados no gráfico são regredidos, isto é, são definidos ou modelados por uma reta que corresponde à menor distância entre cada ponto plotado e a reta.

$Y = \alpha + \beta X$ equação da reta a partir dos dados coletados

$Y' = a + b X'$ equação da reta a partir das estimativas



ERRO OU DESVIO

Haverá sempre alguma diferença entre o valor observado Y e o valor estimado Y' . Essa diferença em estatística é chamada de erro ou desvio:

$$e = Y - Y'$$

O erro indica que:

- As variações de Y não são perfeitamente explicadas pelas variações de X ou;
- Existem outras variáveis das quais Y depende ou;
- Os valores de X e Y são obtidos de uma amostra particular que não é representativa da realidade.



OBJETIVO DE UMA REGRESSÃO

- Reduzir a diferença entre Y (plotado / observado) e Y' (estimado / calculado) ou;
- Tornar mínimos os somatórios dos desvios entre Y e Y' .

$$\sum (Y - Y') = (y_1 - y'_1) + (y_2 - y'_2) + \dots + (y_n - y'_n) = \text{mínimo}$$



OBJETIVO DE UMA REGRESSÃO

A reta de regressão é apenas uma aproximação da realidade.

É um modo útil para indicar a tendência dos dados.

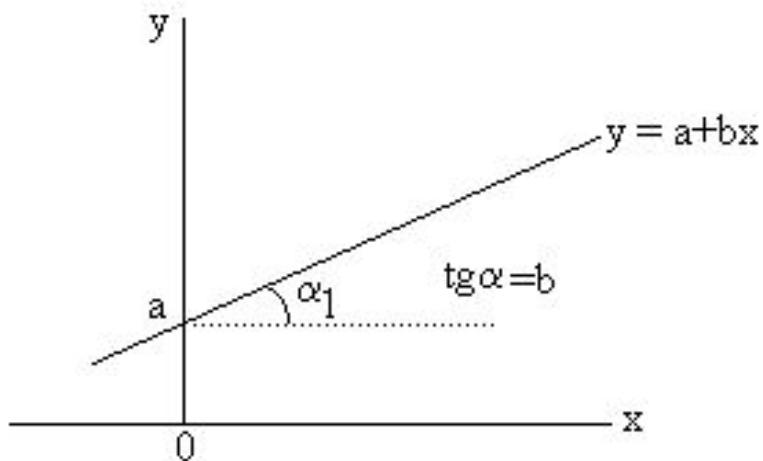
Duas medidas são utilizadas para indicar quanto confiável, útil ou aproximada da realidade é a reta:

- Erro-padrão da estimativa.
- Coeficiente de determinação.



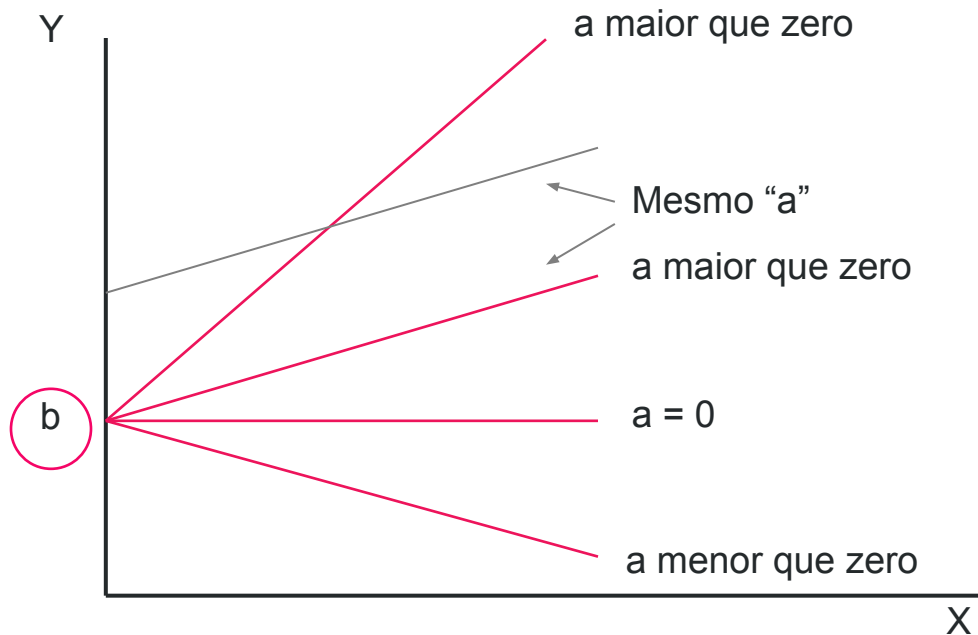
MODELOS DE REGRESSÃO LINEAR

Modelos matemáticos para determinação da relação linear entre variáveis permitem, sob algumas condições, a predição de uma variável em função de outra.



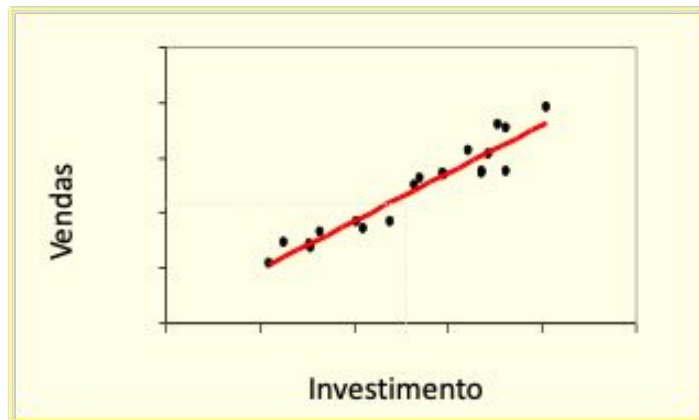
FUNÇÃO LINEAR: A RETA

Função Linear: a reta.



MODELOS DE REGRESSÃO LINEAR

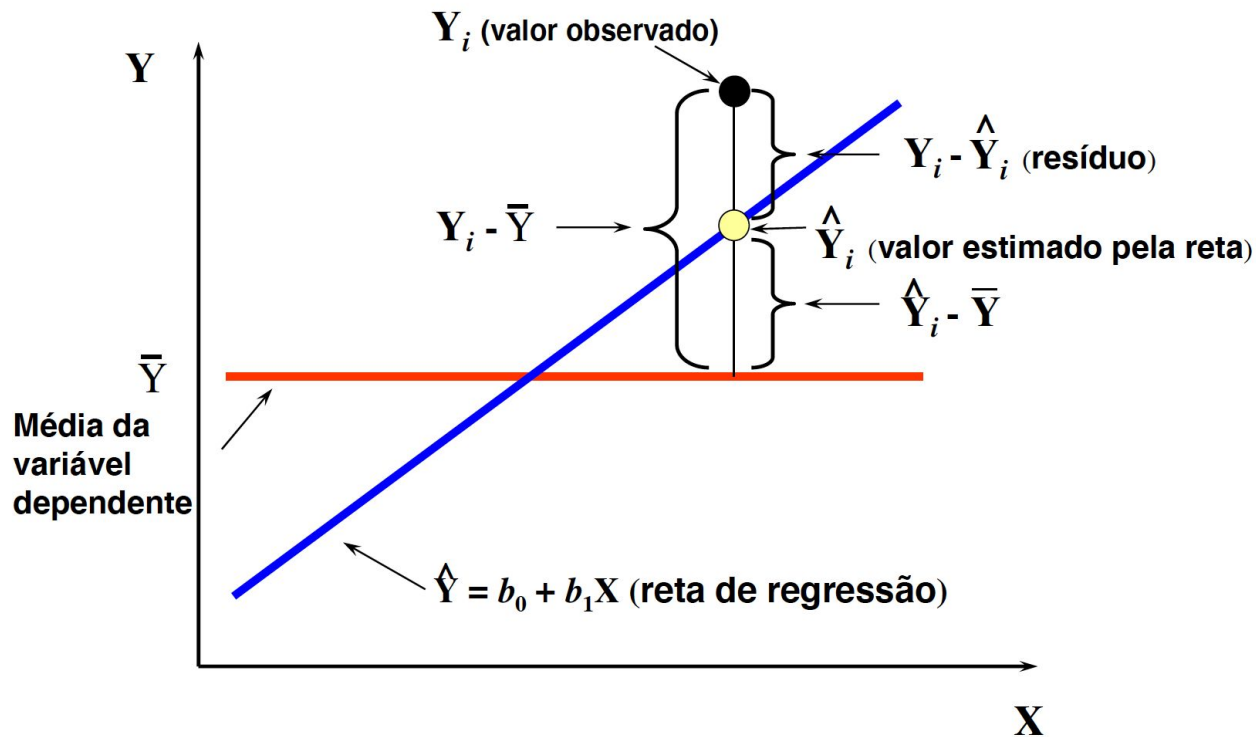
SITUAÇÃO 1: Uma vez verificada a existência de uma relação entre o investimento em treinamento e as vendas de uma empresa, desejamos desenvolver um modelo para estimar a medida de vendas (variável y) a partir da medida dos investimentos em treinamento (variável x).



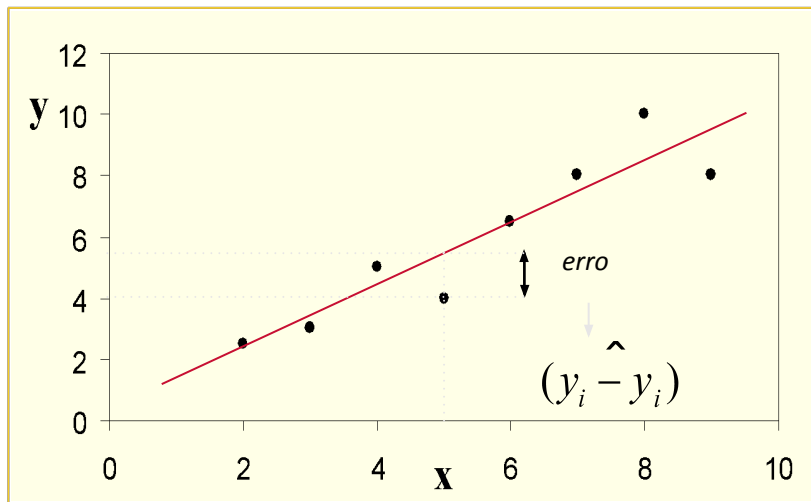
Qual a reta que melhor se ajusta a estes dados?



MODELOS DE REGRESSÃO LINEAR



MÉTODO DOS MÍNIMOS QUADRADOS



O objetivo é minimizar a soma do quadrado dos erros:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Obtendo os valores de b e a que minimizam a equação acima.



MÉTODO DOS MÍNIMOS QUADRADOS

A equação que descreve a relação entre as duas variáveis é:

$$y = \alpha x + b + \varepsilon$$

Podemos utilizar a reta de regressão para estimar os valores de:

$$\hat{y} = a \cdot x + b$$

$$a = \frac{\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i \right) / n}{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2 / n}$$

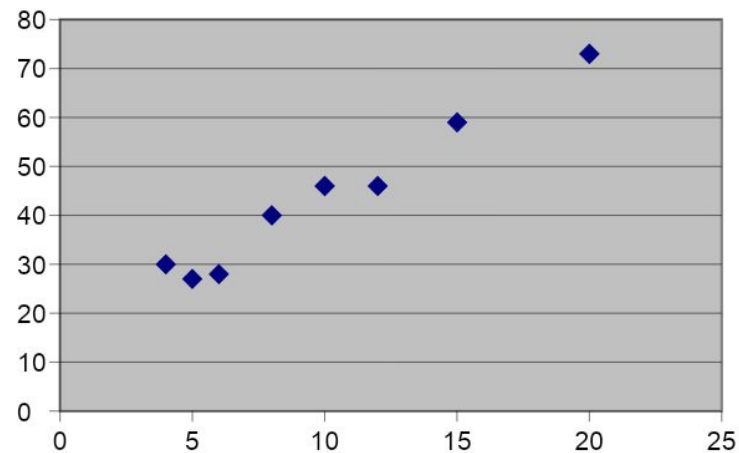
$$b = \bar{y} - a \cdot \bar{x}$$



ESTIMATIVA DOS PARÂMETROS

x_i	y_i
5	27
10	46
20	73
8	40
4	30
6	28
12	46
15	59

$$\hat{y} = b + a \cdot x$$

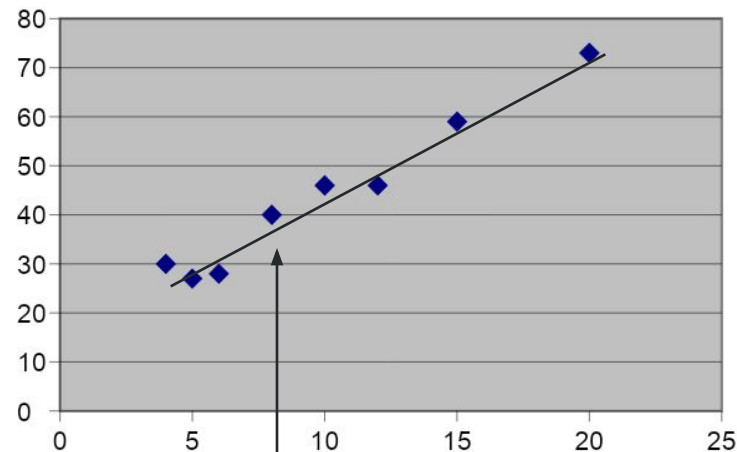


ESTIMATIVA DOS PARÂMETROS

x_i	y_i
5	27
10	46
20	73
8	40
4	30
6	28
12	46
15	59

$$\hat{y} = b + a \cdot x$$

$$\hat{y} = 14,577 + 2,095 \cdot x$$



QUALIDADE DO AJUSTE NA REGRESSÃO

Coeficiente de Determinação

Quando fazemos uma regressão linear, os valores observados (x_i, y_i) estão espalhados ao redor da reta de regressão. Quanto menor for este espalhamento, melhor a reta de regressão representa o conjunto de valores observados. A variância amostral total, como estimador do espalhamento, pode ser decomposta da seguinte forma:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

y_i = Valores observados

\bar{y} = Valor médio

\hat{y}_i = Valores estimados



QUALIDADE DO AJUSTE NA REGRESSÃO

Análise de Variância

A variabilidade total observada na variável dependente está dividida em 2 componentes:

$\hat{Y}_i - Y_i = \text{resíduo da regressão}$

$\hat{Y}_i - \bar{Y}_i = \text{distância da média dos } Y's$

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

*Soma
dos
quadrados
total
SQT ou SST*

*Soma
dos
resíduos
total
SQE ou SSE*

*Soma
dos
quadrados
regressão
SQR ou SSR*

Onde:

SST = total sum of squares.

SSR = sum of squares due to regression.

SSE = sum of squares due to error.



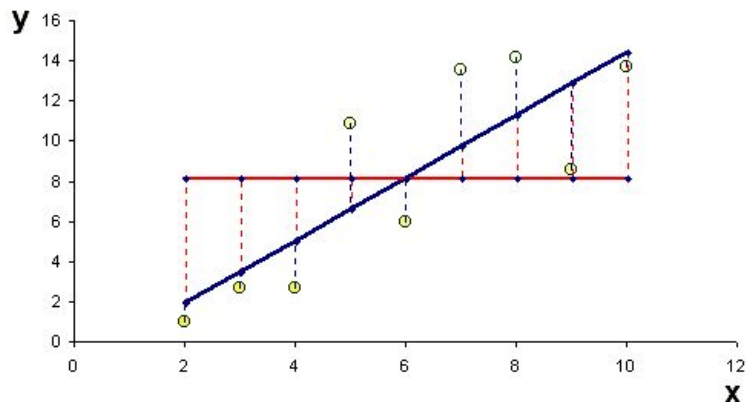
QUALIDADE DO AJUSTE NA REGRESSÃO

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

*Soma
dos
quadrados
total
SQT ou SST*

*Soma
dos
resíduos
total
SQE ou SSE*

*Soma
dos
quadrados
regressão
SQR ou SSR*



$$R^2 = \frac{SSR}{SST} = r^2$$

Onde r é o coeficiente de correlação de Pearson.



<i>Estatística de regressão</i>	
R múltiplo	0.74
R-Quadrado	0.55
R-quadrado ajustado	0.53
Erro padrão	123.27
Observações	25

ANOVA

	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
Regressão	1	428912.3522	428912.3522	28.22781	0.00
Resíduo	23	349477.4932	15194.67362		
Total	24	778389.8455			

	<i>Coefficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
Interseção	194.93	34.13	5.71	0.00	124.33	265.52
VendasCartão (X)	2.11	0.40	5.31	0.00	1.29	2.93



Resultados do Excel

Estatística de regressão	
R múltiplo	0.74
R-Quadrado	0.55
R-quadrado ajustado	0.53
Erro padrão	123.27
Observações	25

A relação linear entre as duas variáveis é medida pelo coeficiente de correlação

R-quadrado da regressão, que mede a proporção da variabilidade em Y que é explicada por X. É uma função direta da correlação entre as variáveis

é uma medida semelhante ao R-quadrado mas que, ao contrário deste, não aumenta com a inclusão de variáveis independentes não significativas

Erro padrão: mede a dispersão dos valores observados em relação a equação da reta

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

Soma de
Soma de
Soma de
Quadrados Total
Quadrados Residual
Quadrados da Regressão

ANOVA

	gl	SQ	MQ	F	F de significação
Regressão	1	428912.3522	428912.3522	28.22781	0.00
Resíduo	23	349477.4932	15194.67362		
Total	24	778389.8455			

A estatística F serve para testar quanto o modelo de regressão ajusta os dados. Se a probabilidade associada com F é pequena, a hipótese que $R^2_{pop} = 0$ é rejeitada.

	Coefficientes	Erro padrão	Stat t (3)	valor-P (4)	95% (5) inferiores	95% superiores
(1) Interseção	194.93	34.13	5.71	0.00	124.33	265.52
(2) VendasCartão (X)	2.11	0.40	5.31	0.00	1.29	2.93

(1) Parâmetro B_0 (intercepto)

(2) Parâmetro B_1 (inclinação da reta)

(3) Teste de hipóteses dos parâmetros B_0 e B_1

(4) Nível descritivo do teste de hipóteses (3)

(5) Intervalo de confiança da estimativa do parâmetro

EQUAÇÃO DA REGRESSÃO LINEAR SIMPLES

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Interceptor

Ponto em que a reta
de regressão toca o
eixo Y.

Coefficiente
da variável
independente.

Erro, uma variável
aleatória não
observável.



EQUAÇÃO DA REGRESSÃO LINEAR MÚLTIPLA

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + + \varepsilon$$

Interceptor

Ponto em que a reta de regressão toca o eixo Y.

Componentes
Determinísticos.

Erro, uma variável
aleatória não
observável.



HIPÓTESES ASSUMIDAS PELO MODELO

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

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.



ROOT MEAN SQUARE ERROR (RMSE)

- A medida de erro normalmente utilizada para avaliar a qualidade do ajuste de um modelo é a chamada RAIZ DO ERRO MÉDIO QUADRÁTICO.
- Ela é a raiz do erro médio quadrático da diferença entre a predição e o valor real.
- Podemos pensar nela como sendo uma medida análoga ao desvio-padrão.
- A medida RMSE tem a mesma unidade que os valores de y .
- RMSE é uma boa medida, porque geralmente ela representa explicitamente o que vários métodos tendem a minimizar.



MEAN ABSOLUTE ERROR (MAE)

Nas estatísticas, o erro absoluto médio (MAE) é uma medida de erros entre observações emparelhadas que expressam o mesmo fenômeno. Exemplos de Y versus X incluem comparações de tempo previsto versus observado, tempo subsequente versus tempo inicial e uma técnica de medição versus uma técnica alternativa de medição.



ERRO ABSOLUTO RELATIVO (RAE)

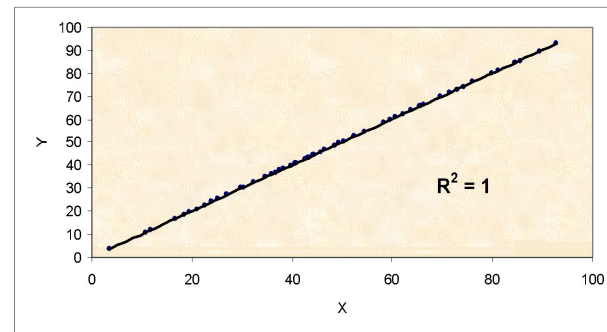
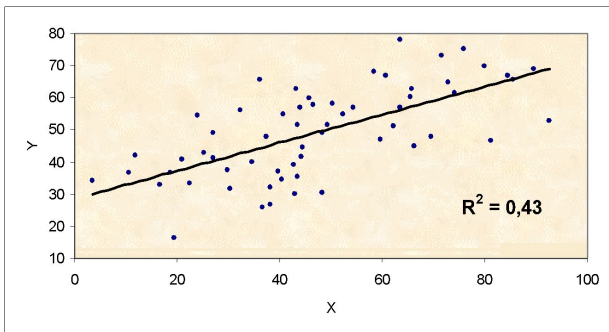
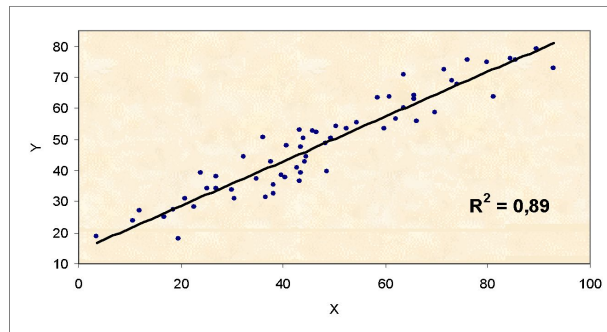
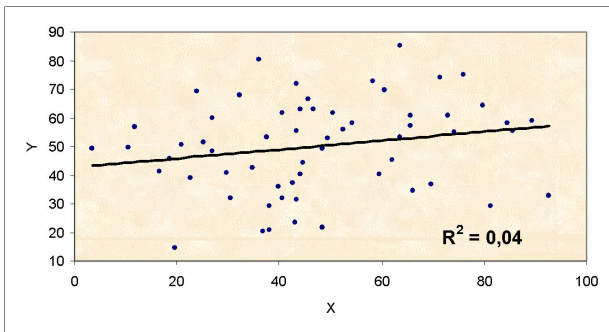
O erro absoluto relativo (RAE) é uma maneira de medir o desempenho de um modelo preditivo usada principalmente em aprendizado de máquina, mineração de dados e gerenciamento de operações. O RAE não deve ser confundido com erro relativo, que é uma medida geral de precisão ou exatidão para instrumentos, como relógios, réguas ou balanças.

O erro absoluto relativo é expresso como uma razão, comparando um erro médio (residual) com os erros produzidos por um modelo trivial ou ingênuo. Um modelo razoável (que produz resultados melhores que um modelo trivial) resultará em uma proporção menor que um.



QUALIDADE DO AJUSTE NA REGRESSÃO

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



EXEMPLO 1

$$\hat{y}_i = 14,577 + 2,905 \cdot x_i$$

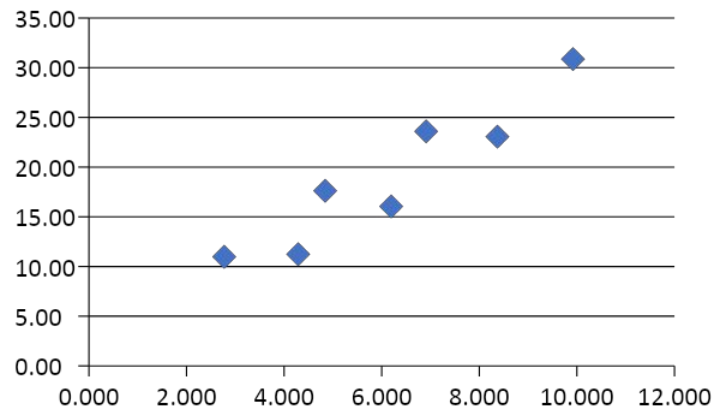
x_i	y_i	\hat{y}_i	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
5	27	29	276,39	210,92	4
10	46	44	5,64	0,000004	6
20	73	73	862,89	844,02	0
8	40	38	13,14	33,73	5
4	30	26	185,64	303,74	14
6	28	32	244,14	134,98	16
12	46	49	5,64	33,78	12
15	59	58	236,39	211,03	1
$\bar{y} = 43,625$			$\Sigma = 1830$	$\Sigma = 1772$	$\Sigma = 58$

$$R^2 = \frac{SSR}{SST} = \frac{1772}{1830} = 0,9683$$

$$SST = SSR + SSE$$

EXEMPLO

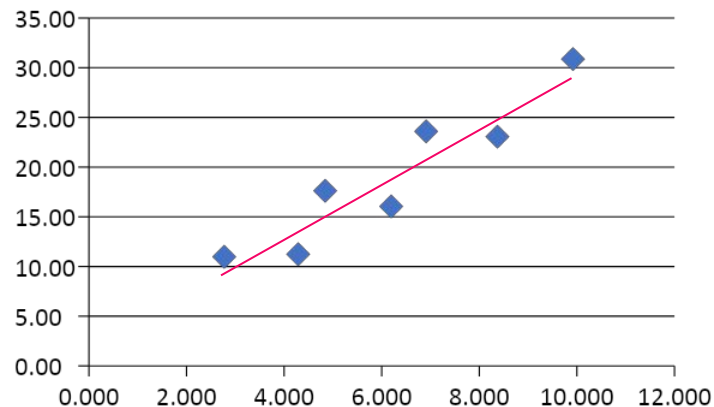
x_i	y_i
2,772	10,98
6,193	16,05
9,917	30,87
4,841	17,61
6,910	23,59
8,372	23,07
4,290	11,24



EXEMPLO

x_i	y_i
2,772	10,98
6,193	16,05
9,917	30,87
4,841	17,61
6,910	23,59
8,372	23,07
4,290	11,24

$$y = 2,769 x + 1,935$$



EXEMPLO

$$\hat{y}_i = 2,769 + 1,935 \cdot x_i$$

x_i	y_i	\hat{y}_i	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
2,772	10,98	9,61	65,26	89,28	1,878
6,193	16,05	19,08	9,05	0,00	9,185
9,917	30,87	29,39	139,51	106,76	2,188
4,841	17,61	15,34	2,10	13,85	5,164
6,910	23,59	21,07	20,53	4,03	6,372
8,372	23,07	25,11	16,09	36,66	4,175
4,290	11,24	13,81	61,13	27,52	6,616
2,772	10,98	9,61	65,26	89,28	1,878
$\bar{y} = 19,06$			$\Sigma = 313,68$	$\Sigma = 278,10$	$\Sigma = 35,58$

$$R^2 = \frac{SSR}{SST} = \frac{278,10}{313,68} = 0,8866$$

$$SST = SSR + SSE$$



MEDIDAS DE DISPERSÃO DO AJUSTE

- Estimativa de s^2 .

O erro médio quadrático (MSE) fornece uma estimativa de s^2 . Usamos a notação s^2 :

$$s^2 = \text{MSE} = \text{SSE}/(n - 2)$$

Onde:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b - a \cdot x_i)^2$$


- Estimativa de s (erro-padrão):

$$s = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{n - 2}}$$



MEDIDAS DE DISPERSÃO DO AJUSTE

$$\hat{y}_i = 2,769 \cdot x_i + 1,935$$



x_i	y_i	\hat{y}_i	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
2,772	10,98	9,61	65,26	89,28	1,878
6,193	16,05	19,08	9,05	0,00	9,185
9,917	30,87	29,39	139,51	106,76	2,188
4,841	17,61	15,34	2,10	13,85	5,164
6,910	23,59	21,07	20,53	4,03	6,372
8,372	23,07	25,11	16,09	36,66	4,175
4,290	11,24	13,81	61,13	27,52	6,616
2,772	10,98	9,61	65,26	89,28	1,878

$$\bar{y} = 19,06$$

$$\Sigma = 313,68$$

$$\Sigma = 278,10$$

$$\Sigma = 35,58$$

$$SST =$$

$$SSR$$

$$+$$

$$SSE$$

$$s^2 = MSE = \frac{SSE}{(N-2)} = \frac{35,58}{5} = 7,116 \Rightarrow s = \sqrt{7,116} = 2,668$$



MEDIDAS DE DISPERSÃO DO AJUSTE

É uma extensão de modelos de regressão linear simples, visto que utiliza mais de uma variável explicativa.

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \dots + e_i$$

EXEMPLO 4:

Dados do laboratório sobre a influência da temperatura média, ou seja, das estações do ano, em relação a um remédio desenvolvido para apoio no combate à gripe.

Vendas no Trimestre (10.000 un)	Despesas com Propaganda (\$ 10.000)	Média da Temperatura (°C)
Vendas	Propaganda	Temperatura
25	9	13
10	7	22
8	4	24
25	12	14
20	9	18
12	6	20
13	5	22
15	6	14
18	8	17



MEDIDAS DE DISPERSÃO DO AJUSTE

<i>Estatística de regressão</i>	
R múltiplo	0,949676887
Quadrado de R	0,901886189
Quadrado de R ajustado	0,869181586
Erro-padrão	2,24259699
Observações	9

ANOVA

	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significância</i>
Regressão	2	277,380108	138,690054	27,57673509	0,000944475
Residual	6	30,17544757	5,029241262		
Total	8	307,5555556			

	<i>Coeficientes</i>	<i>Erro-padrão</i>	<i>Stat t</i>	<i>valor P</i>	<i>95% inferior</i>
Interceptar	19,12813299	7,506193252	2,548313419	0,043589583	0,761139766
Propaganda	1,378005115	0,451520517	3,051921369	0,022456485	0,27317421
Temperatura	-0,718414322	0,263548007	-2,725933429	0,03437098	-1,363293064

$$\text{Vendas} = 19,1281 + 1,378 \text{ Propaganda} - 0,7184 \text{ Temperatura}$$



TESTES DE SIGNIFICÂNCIA DO AJUSTE

Teste F do ajuste

Hipótese $\begin{cases} H_0: \beta_i = 0 \text{ para algum valor de } i \neq 0 \\ H_a: \beta_i \neq 0 \forall i \neq 0 \end{cases}$

Estatística do teste (ANOVA): $F = \text{MSR}/\text{MSE}$

Rejeitar H_0 se valor-p $< \alpha$ ou $F > F_\alpha$

Teste dos coeficientes:

Hipótese: $\begin{cases} H_0: \beta_i = 0 \\ H_1: \beta_i \neq 0 \end{cases}$

Estatística do teste t

Rejeitar H_0 se valor-p $< \alpha$ ou $t > t_\alpha$



MULTICOLINEARIDADE

- Problema que ocorre quando as variáveis explicativas não são independentes.
- Consequência da Multicolinearidade: as estimativas dos parâmetros perdem a confiabilidade.
- Indicações de Multicolinearidade:
 - Resultados obtidos atentam contra o bom senso.
 - Valor-P maior que 0,05.
 - Alta correlação entre as variáveis do modelo.

Ação necessária: **eliminar** alguma variável explicativa e **efetuar uma nova regressão**.



ARMADILHAS DA REGRESSÃO MÚLTIPLA

Vendas (R\$ 1.000)	Propaganda (R\$ 1.000)	Desconto (%)
2562	16,66	0,5
2592	33,34	1,5
2751	40	2
2670	50	4
2880	66,66	4,5
2640	70	5
3110	83,34	5
3120	100	5,5
2811	110	6,5
2838	116,66	8
3258	133,34	8,5
3080	150	9
2925	160	9,5
3495	166,66	10,5
3152	183,34	11
3057	190	12
3424	200	12



ARMADILHAS DA REGRESSÃO MÚLTIPLA

SUMÁRIO DOS RESULTADOS

<i>Estatística de regressão</i>	
R múltiplo	0,803217619
Quadrado de R	0,645158544
Quadrado de R ajustado	0,594466907
Erro-padrão	179,2507745
Observações	17

ANOVA

	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significância</i>
Regressão	2	817866,1201	408933,06	12,72712005	0,000708338
Residual	14	449831,7623	32130,84016		
Total	16	1267697,882			

Vendas (R\$ 1.000)	Propaganda (R\$ 1.000)	Desconto (%)
2562	16,66	0,5
2592	33,34	1,5
2751	40	2
2670	50	4
2880	66,66	4,5
2640	70	5
3110	83,34	5
3120	100	5,5
2811	110	6,5
2838	116,66	8
3258	133,34	8,5
3080	150	9
2925	160	9,5
3495	166,66	10,5
3152	183,34	11
3057	190	12
3424	200	12

	<i>Coefficientes</i>	<i>Erro-padrão</i>	<i>Stat t</i>	<i>valor P</i>	<i>95% inferior</i>	<i>95% superior</i>	<i>inferior 95,0%</i>	<i>superior 95,0%</i>
Interceptar	2540,420837	94,56668608	26,86380312	1,90948E-13	2337,595467	2743,24621	2337,59547	2743,24621
Propaganda (R\$ 1.000)	6,800985526	5,542032663	1,227164461	0,239997399	-5,085492353	18,6874634	-5,0854924	18,6874634
Desconto (%)	-48,1738883	88,96120288	-0,541515703	0,596665179	-238,976692	142,628915	-238,97669	142,628915

Vendas = 2540,42 + 6,8009 propaganda – 48,1733 desconto

ARMADILHAS DA REGRESSÃO MÚLTIPLA

SUMÁRIO DOS RESULTADOS

<i>Estatística de regressão</i>	
R múltiplo	0,803217619
Quadrado de R	0,645158544
Quadrado de R ajustado	0,594466907
Erro-padrão	179,2507745
Observações	17

$$\text{Vendas} = 2540,42 + 6,8009 \text{ propaganda} - 48,1733 \text{ desconto}$$

ANOVA

	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significância</i>
Regressão	2	817866,1201	408933,06	12,72712005	0,000708338
Residual	14	449831,7623	32130,84016		
Total	16	1267697,882			

Vendas (R\$ 1.000)	Propaganda (R\$ 1.000)	Desconto (%)
2562	16,66	0,5
2592	33,34	1,5
2751	40	2
2670	50	4
2880	66,66	4,5
2640	70	5
3110	83,34	5
3120	100	5,5
2811	110	6,5
2838	116,66	8
3258	133,34	8,5
3080	150	9
2925	160	9,5
3495	166,66	10,5
3152	183,34	11
3057	190	12
3424	200	12

	<i>Coefficientes</i>	<i>Erro-padrão</i>	<i>Stat t</i>	<i>valor P</i>	<i>95% inferior</i>	<i>95% superior</i>	<i>nferior 95,0%</i>	<i>superior 95,0%</i>
Interceptar	2540,420837	94,56668608	26,86380312	1,90948E-13	2337,595467	2743,24621	2337,59547	2743,24621
Propaganda (R\$ 1.000)	6,800985526	5,542032663	1,227164461	0,239997399	-5,085492353	18,6874634	-5,0854924	18,6874634
Desconto (%)	-48,1738883	88,96120288	-0,541515703	0,596665179	-238,976692	142,628915	-238,97669	142,628915

	<i>Vendas (R\$ 1.000)</i>	<i>Propaganda (R\$ 1.000)</i>	<i>Desconto (%)</i>
Vendas (R\$ 1.000)	1		
Propaganda (R\$ 1.000)	0,80		
Desconto (%)	0,78	0,99	1

Alta correlação entre as variáveis independentes X_1 e X_2

ARMADILHAS DA REGRESSÃO MÚLTIPLA

SUMÁRIO DOS RESULTADOS

<i>Estatística de regressão</i>	
R múltiplo	0,798577582
Quadrado de R	0,637726155
Quadrado de R ajustado	0,613574565
Erro-padrão	174,9769101
Observações	17

ANOVA

	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significância</i>
Regressão	1	808444,0964	808444,0964	26,40514203	0,000121213
Residual	15	459253,786	30616,91906		
Total	16	1267697,882			

	<i>Coefficientes</i>	<i>Erro-padrão</i>	<i>Stat t</i>	<i>valor P</i>	<i>95% inferior</i>	<i>95% superior</i>	<i>nferior 95,0%</i>	<i>superior 95,0%</i>
Interceptar	2541,513792	92,2909115	27,53807228	2,93964E-14	2344,80037	2738,22721	2344,80037	2738,22721
Propaganda (R\$ 1.000)	3,828484246	0,745045182	5,13859339	0,000121213	2,240458032	5,41651046	2,24045803	5,41651046

Vendas (R\$ 1.000)	Propaganda (R\$ 1.000)	Desconto (%)
2562	16,66	0,5
2592	33,34	1,5
2751	40	2
2670	50	4
2880	66,66	4,5
2640	70	5
3110	83,34	5
3120	100	5,5
2811	110	6,5
2838	116,66	8
3258	133,34	8,5
3080	150	9
2925	160	9,5
3495	166,66	10,5
3152	183,34	11
3057	190	12
3424	200	12

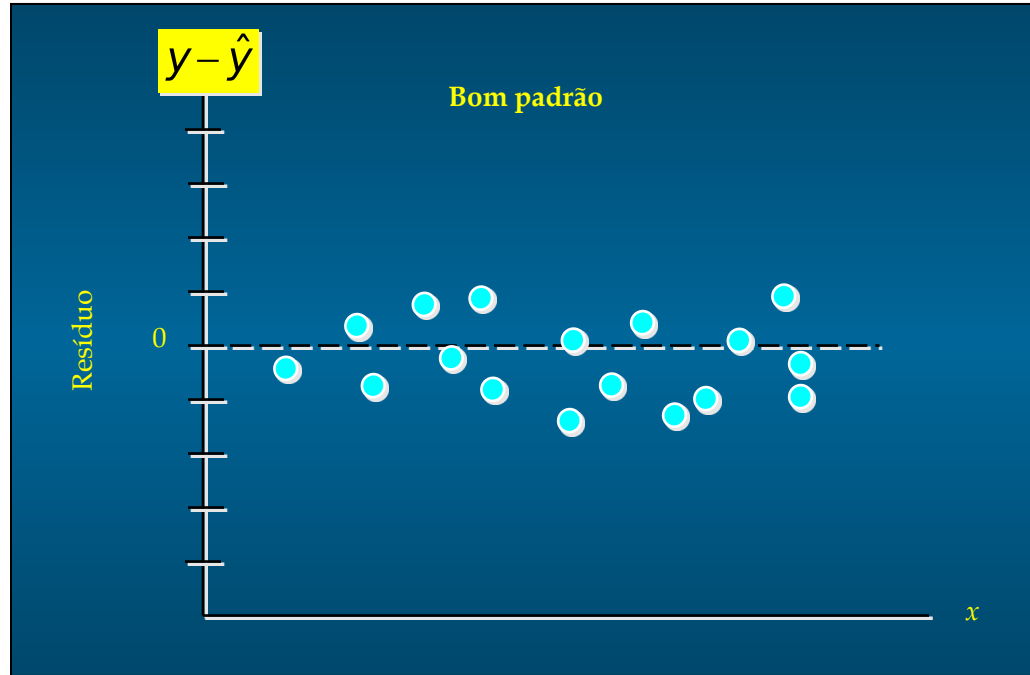
$$\text{Vendas} = 2541,51 + 3,8284 \text{ propaganda}$$

ARMADILHAS DA REGRESSÃO MÚLTIPLA

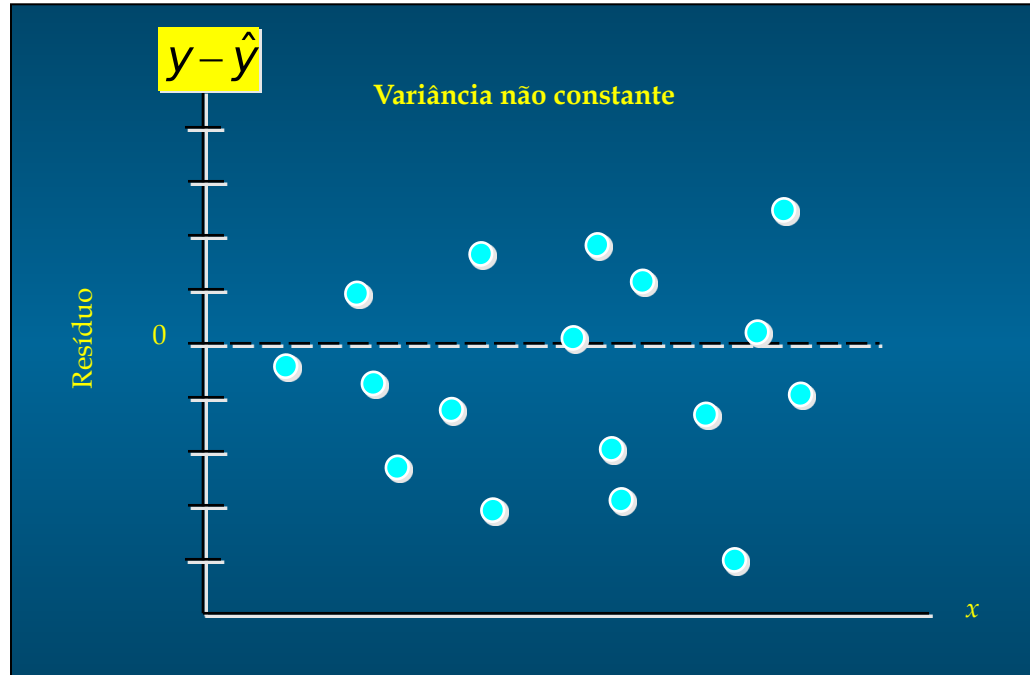
- Resíduo é a diferença entre os valores reais (da amostra) e os valores estimados pelo modelo.
- Deve ser feito um gráfico do resíduo para ser analisado.
- Resultados mais imediatos:
 - Necessidade de termos de ordem superior.
 - Variável significativa não presente.
 - Identificação de outliers.
- O resíduo deve ter média 0 e ser uniformemente distribuído em torno do zero, sem evidência de nenhuma estrutura.



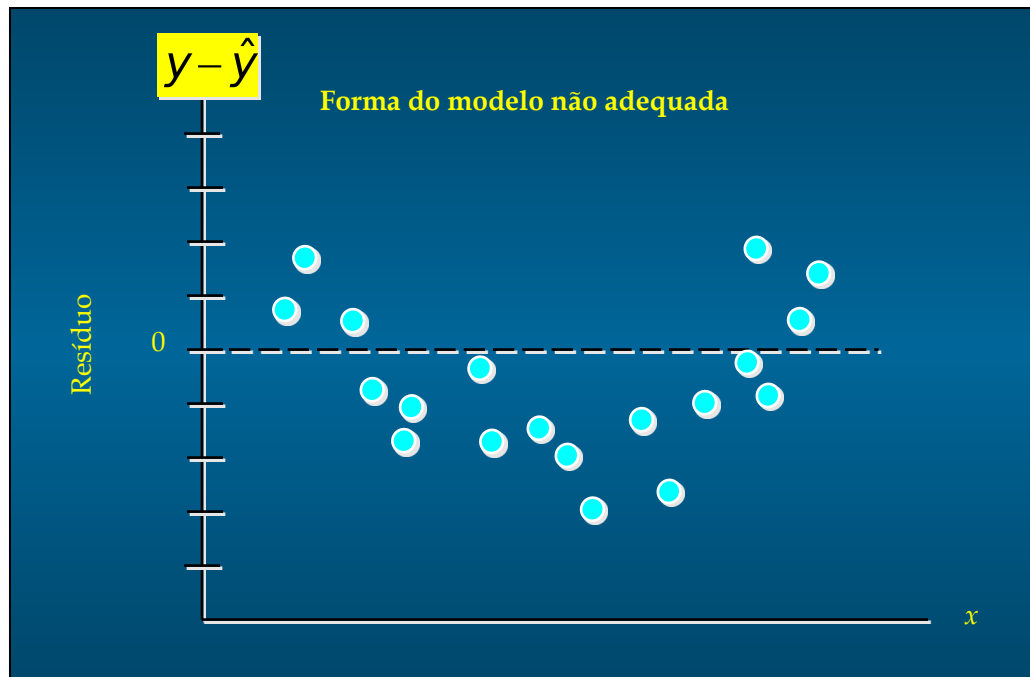
ARMADILHAS DA REGRESSÃO: RESÍDUOS



ARMADILHAS DA REGRESSÃO: RESÍDUOS



ARMADILHAS DA REGRESSÃO: RESÍDUOS



OBRIGADO

 / andresilvadecarvalho



lattes.cnpq.br/6876528572507972

FIAP

Copyright © 2021 | Professor André Silva de Carvalho

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor



SHIFT

 FIAP

