Regressão Linear Simples

Profa Ana Amélia Benedito Silva

aamelia@usp.br

Correlação e Regressão

Objetivo da aula:
estudar a relação entre duas variáveis quantitativas

- Idade e altura de crianças
- Tempo de prática de esportes e ritmo cardíaco
- Tempo de estudo e nota de provas
- Taxa de desemprego e taxa de mortalidade
- Preço de um artigo e a quantidade procurada
- Renda per capita e índice de analfabetismo de países
- Expectativa de vida e taxa de analfabetismo
- Força muscular e capacidade funcional em pacientes com artrite
- Temperatura ambiente e rendimento de um motor
- Peso e altura
- Número de funcionários por agência e a classificação da agência
- Notas de cálculo e estatística em uma classe
- Altura e classificação de atletas numa prova esportiva

Correlação e Regressão

Investigaremos a presença ou ausência de relação linear sob dois pontos de vista:

a) Quantificando a força dessa relação: correlação.

b) Explicitando a forma dessa relação: regressão.

Representação gráfica de duas variáveis quantitativas: Diagrama de dispersão

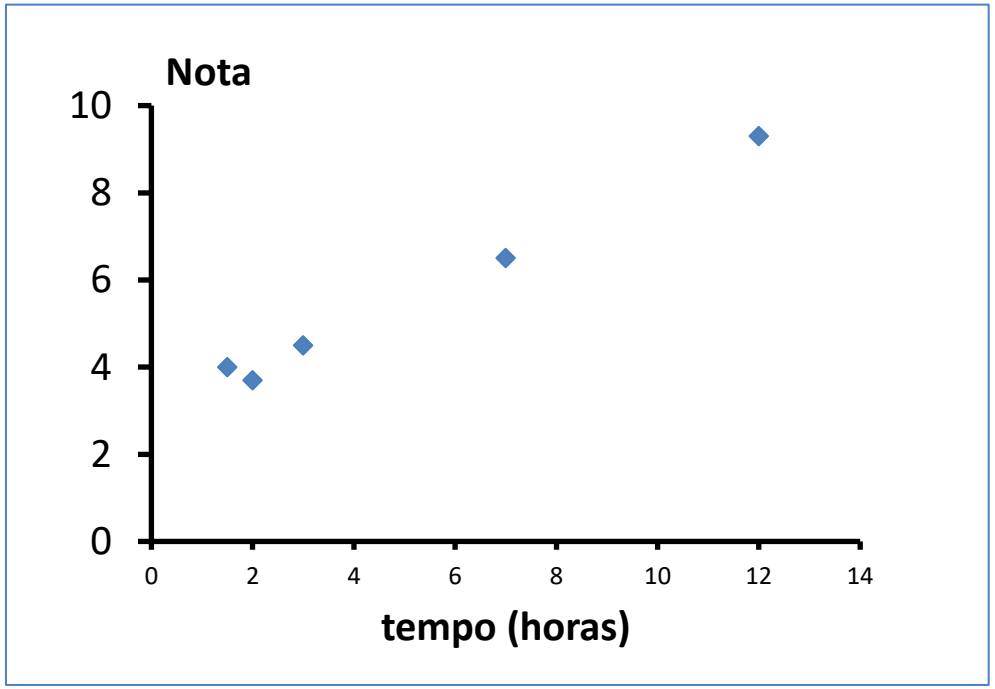
Exemplo 1: nota da prova e tempo de estudo

X : tempo de estudo (em horas)

Y : nota da prova

Pares de observações (X_i , Y_i) para cada estudante

Tempo de estudo (x)	Nota na prova (y)
3	4.5
7	6.5
2	3.7
1.5	4.0
12	9.3



Coeficiente de correlação linear

É uma medida que avalia o quanto a "nuvem de pontos" no diagrama de dispersão aproxima-se de uma reta.

O **coeficiente de correlação linear de Pearson** é dado por:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{(n-1)S_X S_Y},$$

sendo que

$\overline{X}$ e $\overline{Y}$ são as médias amostrais de X e Y, respectiva mente,

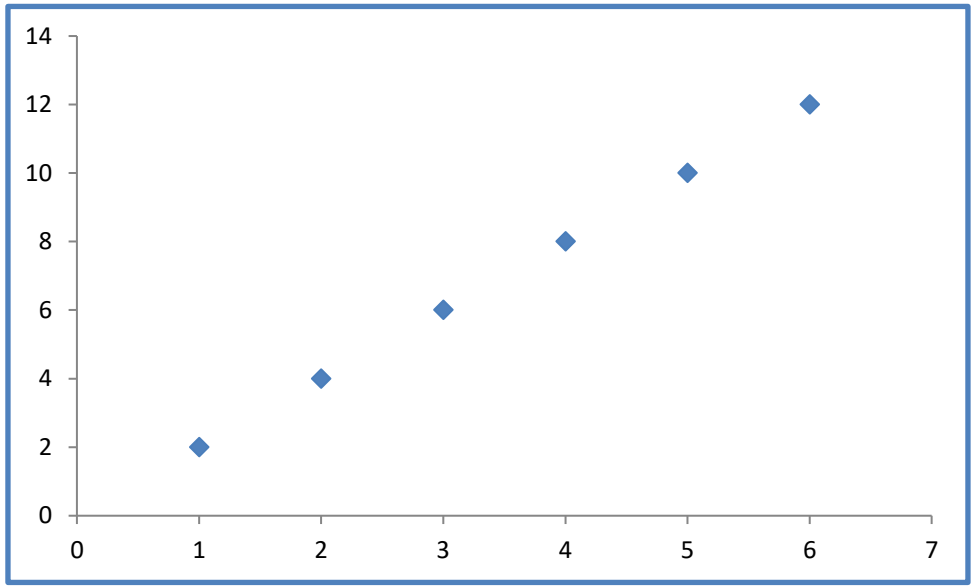
S_X e S_Y são os desvios padrão de X e Y, respectiva mente.

Fórmula alternativa

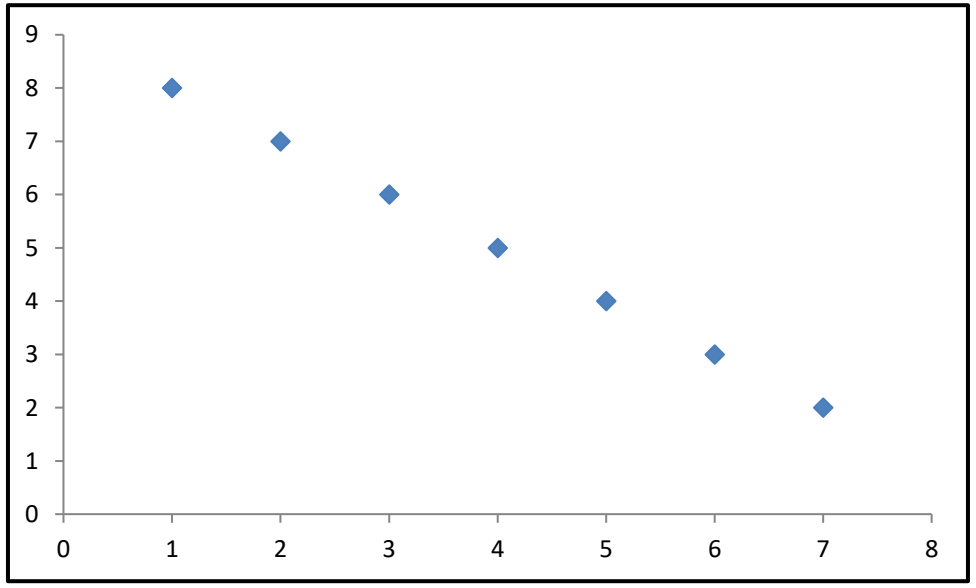
$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{[\sum X^2 - \frac{(\sum X)^2}{n}][\sum Y^2 - \frac{(\sum Y)^2}{n}]}}$$

Coeficiente de correlação (r)

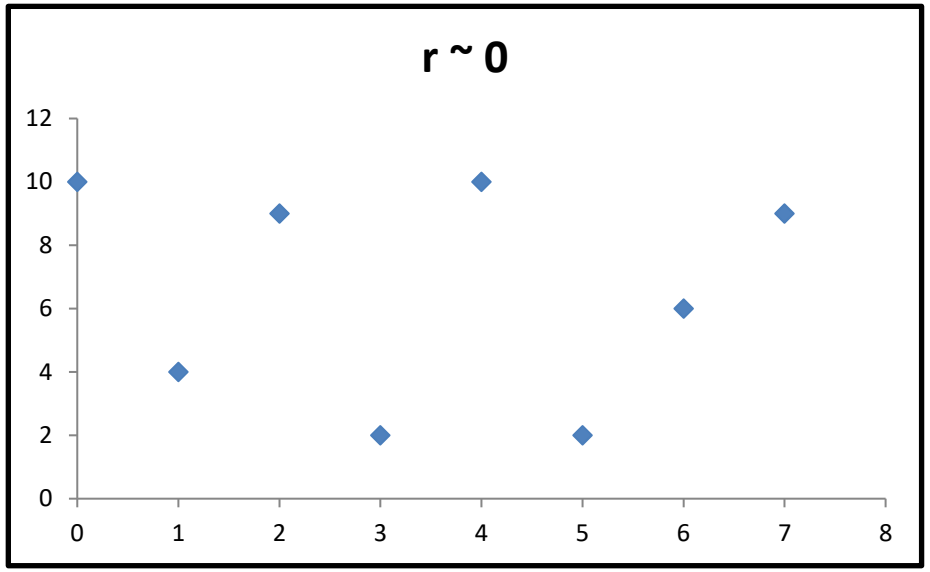
- "mede" a força de associação entre as 2 variáveis
- varia entre -1 e +1 ($-1 \leq r \leq 1$)
- r é uma medida de relação linear; quando a relação não é linear, não é adequado.
- **O coeficiente de correlação não é uma porcentagem!!**

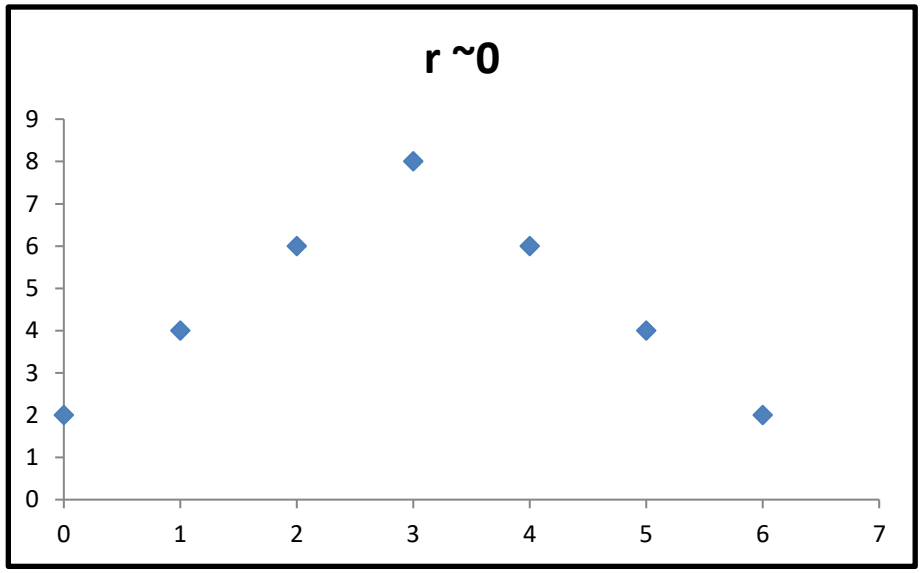


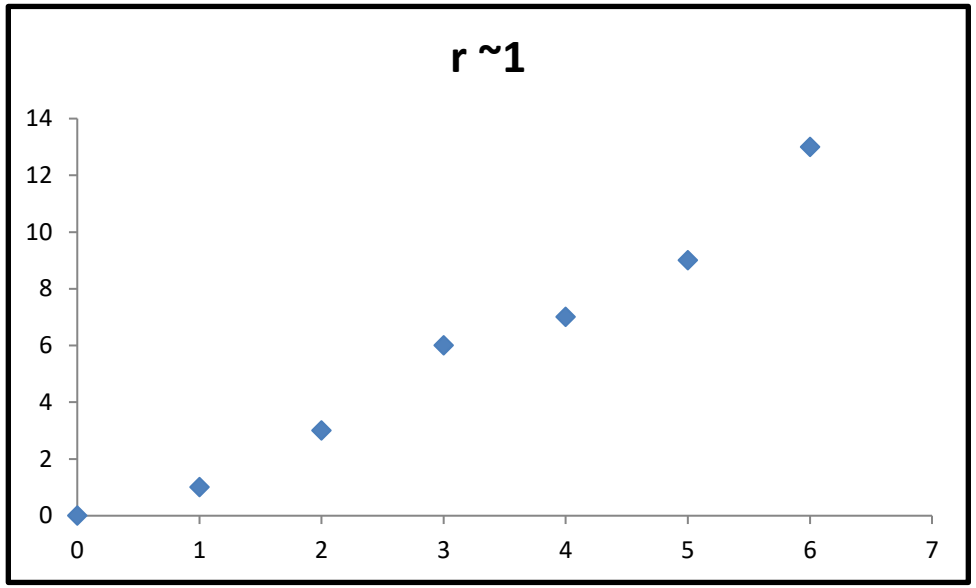
r =1
correlação
linear positiva
e perfeita

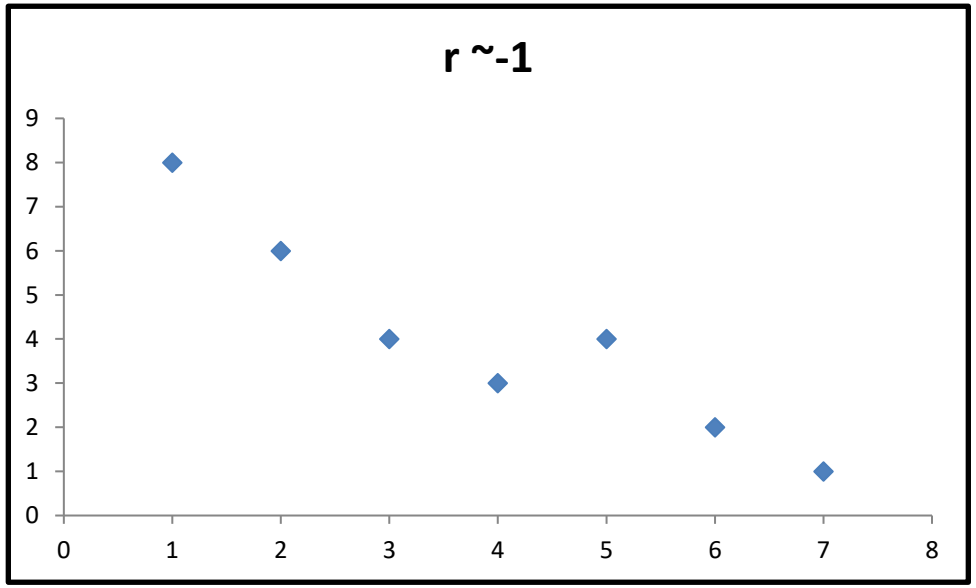


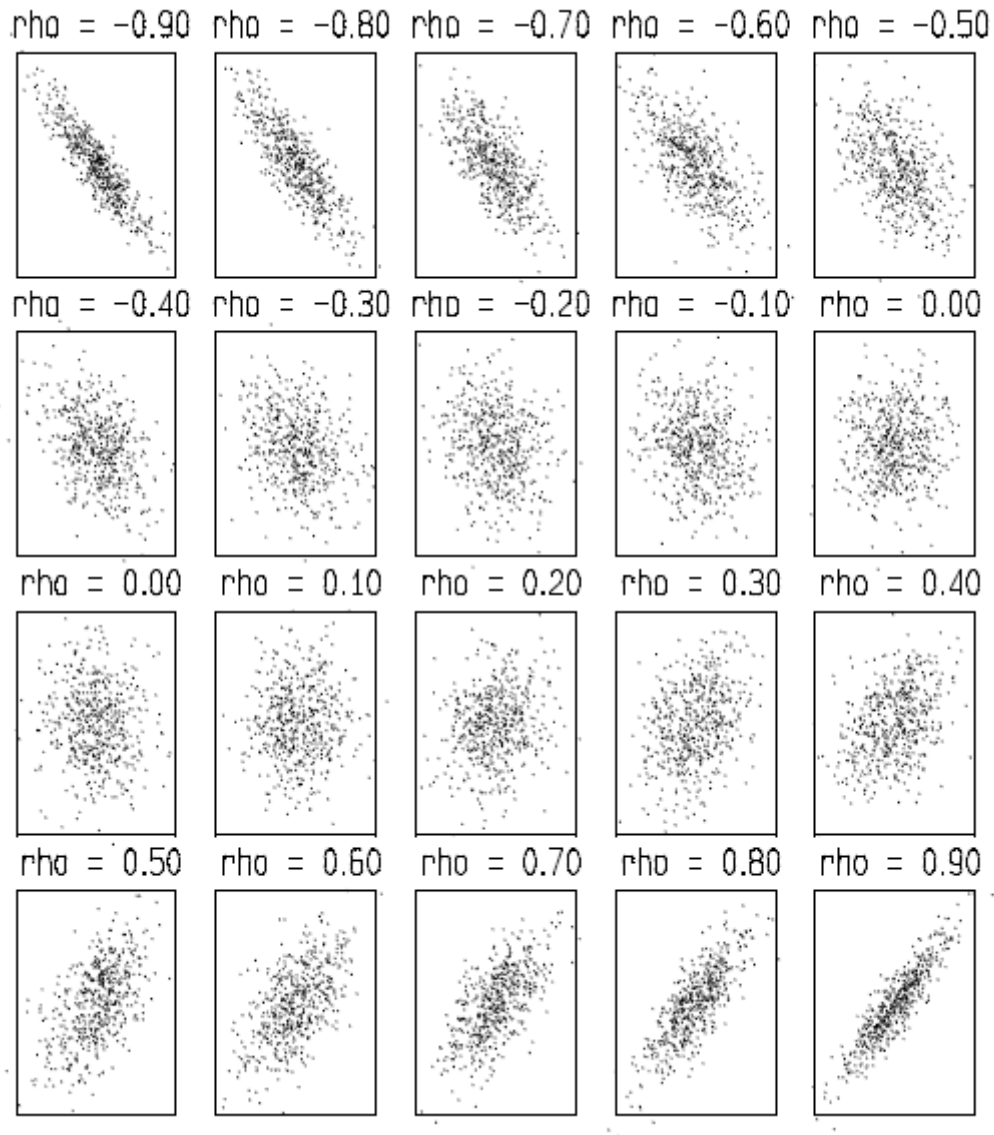
r =-1
correlação
linear negativa
e perfeita


r ~ 0
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7
8


r ~0
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7


r ~1
14
12
10
8
6
4
2
0
0
1
2
3
4
5
6
7

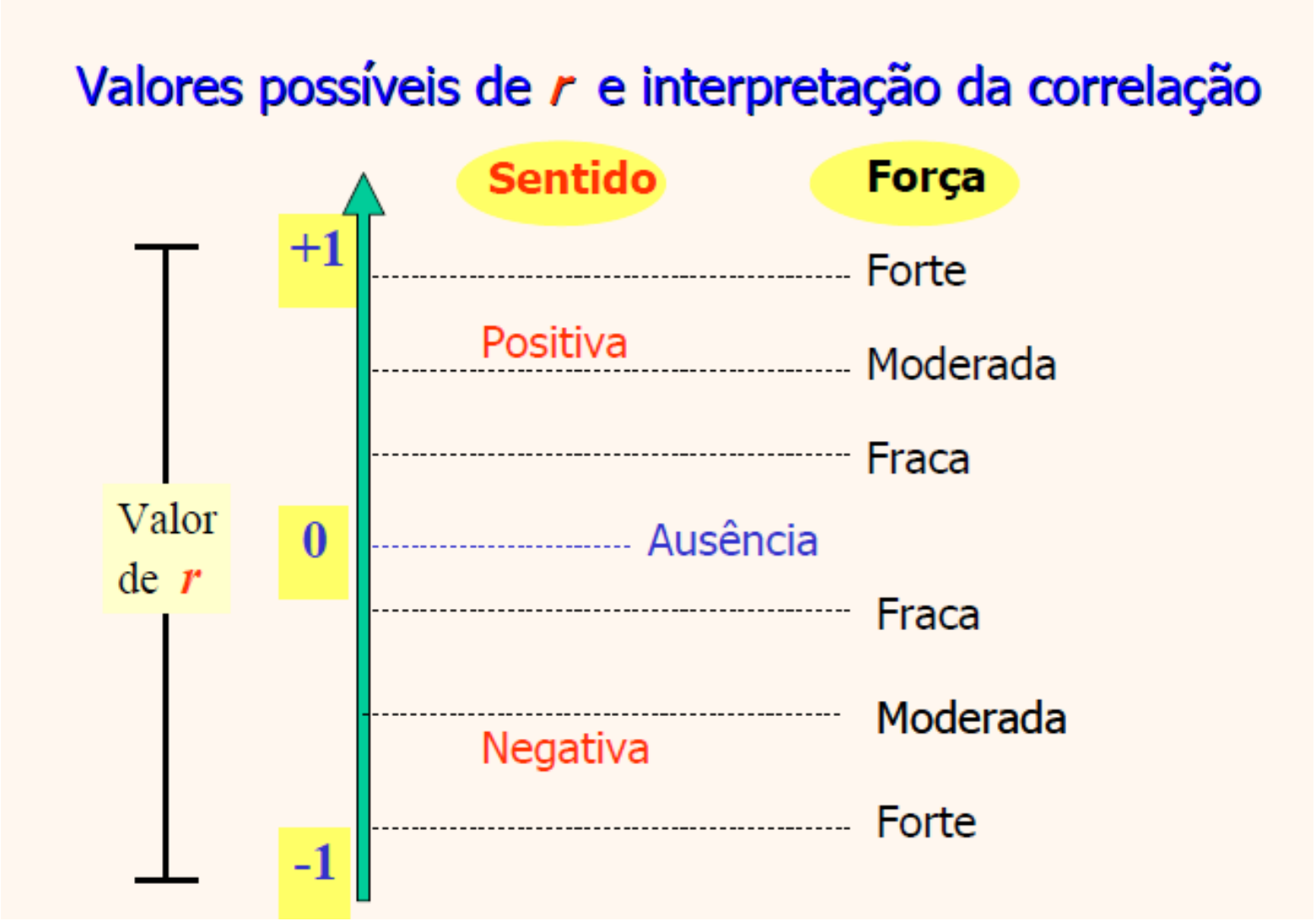

r ~-1
9
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8


rho = -0.90
rho = -0.80
rho = -0.70
rho = -0.60
rho = -0.50
rho = -0.40
rho = -0.30
rho = -0.20
rho = -0.10
rho = 0.00
rho = 0.00
rho = 0.10
rho = 0.20
rho = 0.30
rho = 0.40
rho = 0.50
rho = 0.60
rho = 0.70
rho = 0.80
rho = 0.90

Critérios para avaliar os coeficientes de correlação (em módulo)

$r < 0.25$	: relação baixa ou inexistente
$0.25 \leq r < 0.50$	: relação fraca
$0.50 \leq r < 0.75$	: relação moderada a boa
$r \geq 0.75$	: relação boa a excelente

Outros critérios:



No exemplo:

Tempo (X)	Nota (Y)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$
3,0	4,5	-2,1	-1,1	2,31
7,0	6,5	1,9	0,9	1,71
2,0	3,7	-3,1	-1,9	5,89
1,5	4,0	-3,6	-1,6	5,76
12,0	9,3	6,9	3,7	25,53
25,5	28,0	0	0	41,2

$\bar{X} = 5,1$ $\bar{Y} = 5,6$

$$S_x^2 = \frac{(-2,1)^2 + \ldots + (6,9)^2}{4} = \frac{78,2}{4} = 19,55 \Rightarrow S_x = 4,42$$

$$S_y^2 = \frac{(-1,1)^2 + \ldots + (3,7)^2}{4} = \frac{21,9}{4} = 5,47 \Rightarrow S_y = 2,34$$

Então,

$$r = \frac{41,2}{4 \,.\, 4,42 \,.\, 2,34} = 0,9959$$

Exemplo 2: criminalidade e analfabetismo

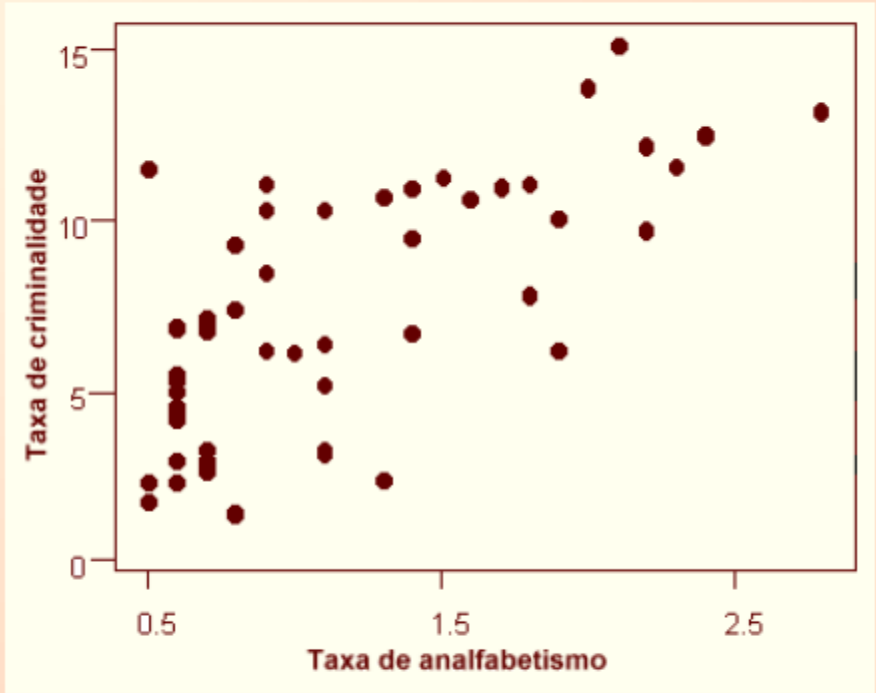
Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: taxa de criminalidade

X: taxa de analfabetismo

Obs.	Estado	Tanalf-70	Exvida-70	Tcrime-75	Obs.	Estado	Tanalf-70	Exvida-70	Tcrime-75
1	Alabama	2.1	69.05	15.1	26	Montana	0.6	70.56	5
2	Alaska	1.5	69.31	11.3	27	Nebraska	0.6	72.6	2.9
3	Arizona	1.8	70.55	7.8	28	Nevada	0.5	69.03	11.5
4	Arkansas	1.9	70.66	10.1	29	New-Hampshire	0.7	71.23	3.3
5	California	1.1	71.71	10.3	30	New-Jersey	1.1	70.93	5.2
6	Colorado	0.7	72.06	6.8	31	New-Mexico	2.2	70.32	9.7
7	Connecticut	1.1	72.48	3.1	32	New-York	1.4	70.55	10.9
8	Delaware	0.9	70.06	6.2	33	North-Carolina	1.8	69.21	11.1
9	Florida	1.3	70.66	10.7	34	North-Dakota	0.8	72.78	1.4
10	Georgia	2	68.54	13.9	35	Ohio	0.8	70.82	7.4
11	Hawaii	1.9	73.6	6.2	36	Oklahoma	1.1	71.42	6.4
12	Idaho	0.6	71.87	5.3	37	Oregon	0.6	72.13	4.2
13	Illinois	0.9	70.14	10.3	38	Pennsylvania	1	70.43	6.1
14	Indiana	0.7	70.88	7.1	39	Rhode-Island	1.3	71.9	2.4
15	Iowa	0.5	72.56	2.3	40	South-Carolina	2.3	67.96	11.6
16	Kansas	0.6	72.58	4.5	41	South-Dakota	0.5	72.08	1.7
17	Kentucky	1.6	70.1	10.6	42	Tennessee	1.7	70.11	11
18	Louisiana	2.8	68.76	13.2	43	Texas	2.2	70.9	12.2
19	Maine	0.7	70.39	2.7	44	Utah	0.6	72.9	4.5
20	Maryland	0.9	70.22	8.5	45	Vermont	0.6	71.64	5.5
21	Massachusetts	1.1	71.83	3.3	46	Virginia	1.4	70.08	9.5
22	Michigan	0.9	70.63	11.1	47	Washington	0.6	71.72	4.3
23	Minnesota	0.6	72.96	2.3	48	West-Virginia	1.4	69.48	6.7
24	Mississippi	2.4	68.09	12.5	49	Wisconsin	0.7	72.48	3
25	Missouri	0.8	70.69	9.3	50	Wyoming	0.6	70.29	6.9

Diagrama de dispersão



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a taxa de criminalidade (Y) tende a aumentar. Nota-se também uma tendência linear.

Cálculo da correlação

$\bar{Y}$= 7,38 (média de Y) e S_Y = 3,692 (desvio padrão de Y)

$\bar{X}$= 1,17 (média de X) e S_x = 0,609 (desvio padrão de X)

$\Sigma X_i Y_i$ = 509,12

Correlação entre X e Y:

$$r = \frac{\sum_{1}^{n} X_i\, Y_i - n\,\bar{X}\,\bar{Y}}{(n-1)\,S_x\,S_y}$$

$$r = \frac{509{,}12 - 50 \cdot 7{,}38 \cdot 1{,}17}{49 \cdot 3{,}692 \cdot 0{,}609} = \frac{77{,}39}{110{,}17} = 0{,}702$$

Exemplo 3: expectativa de vida e analfabetismo

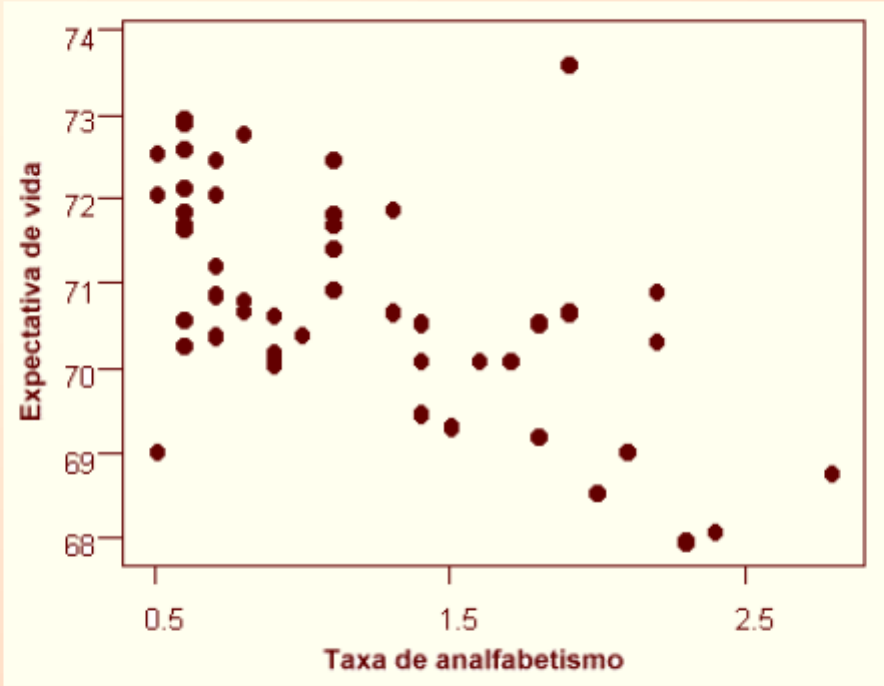
Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: expectativa de vida

X: taxa de analfabetismo

Obs.	Estado	Tanalf-70	Exvida-70	Tcrime-75	Obs.	Estado	Tanalf-70	Exvida-70	Tcrime-75
1	Alabama	2.1	69.05	15.1	26	Montana	0.6	70.56	5
2	Alaska	1.5	69.31	11.3	27	Nebraska	0.6	72.6	2.9
3	Arizona	1.8	70.55	7.8	28	Nevada	0.5	69.03	11.5
4	Arkansas	1.9	70.66	10.1	29	New-Hampshire	0.7	71.23	3.3
5	California	1.1	71.71	10.3	30	New-Jersey	1.1	70.93	5.2
6	Colorado	0.7	72.06	6.8	31	New-Mexico	2.2	70.32	9.7
7	Connecticut	1.1	72.48	3.1	32	New-York	1.4	70.55	10.9
8	Delaware	0.9	70.06	6.2	33	North-Carolina	1.8	69.21	11.1
9	Florida	1.3	70.66	10.7	34	North-Dakota	0.8	72.78	1.4
10	Georgia	2	68.54	13.9	35	Ohio	0.8	70.82	7.4
11	Hawaii	1.9	73.6	6.2	36	Oklahoma	1.1	71.42	6.4
12	Idaho	0.6	71.87	5.3	37	Oregon	0.6	72.13	4.2
13	Illinois	0.9	70.14	10.3	38	Pennsylvania	1	70.43	6.1
14	Indiana	0.7	70.88	7.1	39	Rhode-Island	1.3	71.9	2.4
15	Iowa	0.5	72.56	2.3	40	South-Carolina	2.3	67.96	11.6
16	Kansas	0.6	72.58	4.5	41	South-Dakota	0.5	72.08	1.7
17	Kentucky	1.6	70.1	10.6	42	Tennessee	1.7	70.11	11
18	Louisiana	2.8	68.76	13.2	43	Texas	2.2	70.9	12.2
19	Maine	0.7	70.39	2.7	44	Utah	0.6	72.9	4.5
20	Maryland	0.9	70.22	8.5	45	Vermont	0.6	71.64	5.5
21	Massachusetts	1.1	71.83	3.3	46	Virginia	1.4	70.08	9.5
22	Michigan	0.9	70.63	11.1	47	Washington	0.6	71.72	4.3
23	Minnesota	0.6	72.96	2.3	48	West-Virginia	1.4	69.48	6.7
24	Mississippi	2.4	68.09	12.5	49	Wisconsin	0.7	72.48	3
25	Missouri	0.8	70.69	9.3	50	Wyoming	0.6	70.29	6.9

Diagrama de dispersão



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a expectativa de vida (Y) tende a diminuir. Nota-se também uma tendência linear.

Cálculo da correlação

$\bar{Y}$= 70,88 (média de Y) e S_Y = 1,342 (desvio padrão de Y)

$\bar{X}$= 1,17 (média de X) e S_x = 0,609 (desvio padrão de X)

$\Sigma X_i Y_i$ = 4122,8

Correlação entre X e Y:

$$r = \frac{\sum_{1}^{n} X_i\, Y_i - n\,\bar{X}\,\bar{Y}}{(n-1)\,S_x\,S_y}$$

$$r = \frac{4122{,}8 - 50 \cdot 70{,}88 \cdot 1{,}17}{49 \cdot 1{,}342 \cdot 0{,}609} = \frac{-23{,}68}{40{,}047} = -0{,}59$$

Exercício: A tabela abaixo mostra o Tempo de entrega (*Y*) de dez carregamentos em dias em função da distância rodoviária (*X*) em km.

Pergunta: Existe correlação entre a distância rodoviária percorrida e o tempo de entrega de carregamentos ?

X	Y
215	1,0
480	1,0
325	1,5
550	2,0
920	3,0
670	3,0
825	3,5
1070	4,0
1350	4,5
1215	5,0

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{[\sum X^2 - \frac{(\sum X)^2}{n}][\sum Y^2 - \frac{(\sum Y)^2}{n}]}}$$

X	Y	XY	X^2	Y^2
215	1,0			
480	1,0			
325	1,5			
550	2,0			
920	3,0			
670	3,0			
825	3,5			
1070	4,0			
1350	4,5			
1215	5,0			
Σx=	Σy=	Σxy=	Σx^2=	Σy^2=

X	Y	XY	X^2	Y^2
215	1,0	215	46225	1,00
480	1,0	480	230400	1,00
325	1,5	487,5	105625	2,25
550	2,0	1100	302500	4,00
920	3,0	2760	846400	9,00
670	3,0	2010	448900	9,00
825	3,5	2887,5	680625	12,25
1070	4,0	4280	1144900	16,00
1350	4,5	6075	1822500	20,25
1215	5,0	6075	1476225	25,00
$\Sigma x=7620$	$\Sigma y=28,5$	$\Sigma xy=26370$	$\Sigma x^2=7104300$	$\Sigma y^2=99,75$

r = 0,9489

Coeficiente de correlação populacional

$$\rho = Corr(X,Y) = E\left\{\left(\frac{X-\mu_X}{\sigma_X}\right)\cdot\left(\frac{Y-\mu_Y}{\sigma_Y}\right)\right\}$$

$$\mu_X = E(X) \qquad \sigma_X = \sqrt{V(X)}$$

$$\mu_Y = E(Y) \qquad \sigma_Y = \sqrt{V(Y)}$$

Inferência sobre ρ

- Dada uma amostra aleatória simples (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) do par de variáveis aleatórias (X, Y), o coeficiente r pode ser considerado uma *estimativa* do verdadeiro e desconhecido coeficiente ρ

Teste de significância de ρ

- H_0: $\rho = 0$ (as variáveis X e Y são *não correlacionadas*)
- H_1: $\rho \neq 0$ (as variáveis X e Y são *correlacionadas*)

  (pode também ser unilateral)

- No exemplo 2, r = 0,702 para n=50
- Na Tabela para o coeficiente de Correlação de Pearson, $r_{crítico}$ = 0,279
- Portanto, rejeito H_0 a um nível de significância de 5%.

- No exemplo 3, r = -0,59 para n = 50
- Na Tabela para o coeficiente de Correlação de Pearson, $r_{crítico}$ = 0,279
- Portanto, rejeito H_0 a um nível de significância de 5%.

Regressão linear simples

Variável independente, *X*		Variável dependente, *Y*
Temperatura do forno (^{0}C)	→	Resistência mecânica da cerâmica (MPa)
Quantidade de aditivo (%)	→	Octanagem da gasolina
Renda (R$)	→	Consumo (R$)
Memória RAM do computador (Gb)	→	Tempo de resposta do sistema (s)
Área construída do imóvel (m^2)	→	Preço do imóvel (R$)

Considere um experimento em que se analisa a octanagem da gasolina (Y) em função da adição de um novo aditivo (X). Para isso foram realizados ensaios com os percentuais de 1, 2, 3, 4, 5 e 6% de aditivos.

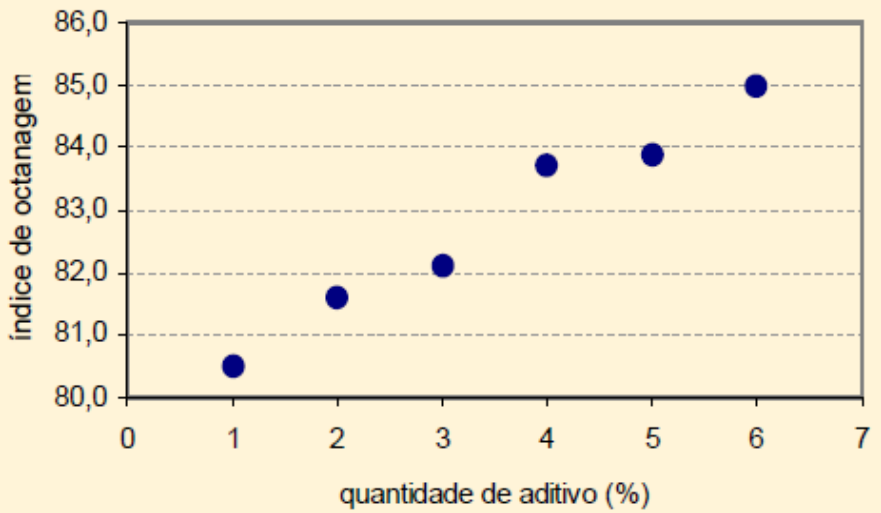
É razoável supor uma relação aproximadamente linear entre X e Y para os níveis de aditivo ensaiados.

Os pontos não são exatamente sobre uma reta provavelmente pela existência de fatores não controláveis no processo.

- X = % de aditivo
- Y = Índice de octanagem da gasolina

Resultados de n = 6 ensaios experimentais:

X	Y
1	80,5
2	81,6
3	82,1
4	83,7
5	83,9
6	85,0



Modelo de regressão linear simples

- Vamos supor que o valor esperado de Y varie com X de acordo com uma equação de 1º grau, onde α e β são os parâmetros.

$$E(Y) = \alpha + \beta X$$

- Seja um conjunto de observações (x_1, y_1), (x_2, y_2). (x_3, y_3)..... (x_n, y_n).

- O modelo de regressão linear é dado por

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

onde

- Y_i é a variável aleatória associada à i-ésima observação de Y
- ε_i é o erro aleatório da i-ésima observação, isto é, o efeito de uma infinidade de fatores que estão afetando a observação de Y de forma aleatória.

Método dos mínimos quadrados

- Para a construção do modelo descrito precisamos obter estimativas para α e β, a partir de um conjunto de observações (x_1, y_1), (x_2, y_2), ...,(x_n, y_n)
- Queremos encontrar a reta que passe o mais próximo possível dos pontos observados.
- O método mais usual para estimar os parâmetros do modelo é o **método dos mínimos quadrados** que consiste em fazer com que a soma dos erros quadráticos seja a menor possível.

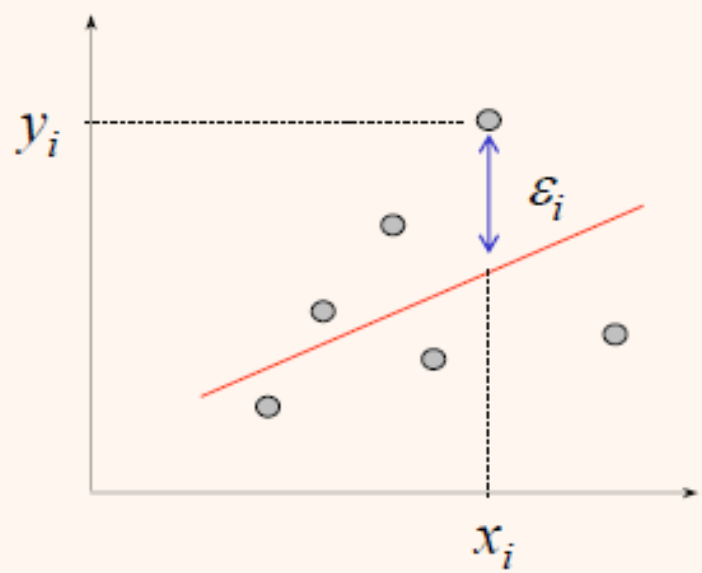
Método dos mínimos quadrados para estimar α e β

- Minimizar em relação a α e β :

$$S = \sum \varepsilon_i^2 = \sum \{Y_i - (\alpha + \beta x_i)\}^2$$

$$\frac{\partial S}{\partial \alpha} = 0$$

$$\frac{\partial S}{\partial \beta} = 0$$



Método dos mínimos quadrados para estimar α e β

- Resultado das derivadas parciais:

Estimativa de β: $$b = \frac{n.\sum(x_i y_i) - (\sum x_i)\cdot(\sum y_i)}{n.\sum x_i^2 - (\sum x_i)^2}$$

Estimativa de α : $$a = \frac{\sum y_i - b\sum x_i}{n}$$

Reta de regressão construída com os dados:

$$\hat{y} = a + bx$$

Exemplo numérico

Dados			Cálculos intermediários	
Ensaio (i)	x_i (aditivo)	y_i (octanagem)	x_i^2	$x_i y_i$
1	1	80,5	1	80,5
2	2	81,6	4	163,2
3	3	82,1	9	246,3
4	4	83,7	16	334,8
5	5	83,9	25	419,5
6	6	85,0	36	510,0
Soma	21	496,8	91	1754,3

$$b = \frac{6.(1754,3) - (21).(496,8)}{6.(91) - (21)^2} = 0,886$$

$$a = \frac{496,8 - (0,886).(21)}{6} = 79,7$$

$$\hat{y} = 79,7 + 0,886x$$ ← Equação da reta

Interpretação do modelo

- O coeficiente b fornece uma estimativa da variação esperada de Y, a partir da variação de uma unidade em X.
- O sinal deste coeficiente indica o sentido da variação
- No exemplo, a cada 1% a mais do aditivo, esperamos um aumento de 0,886 no índice de octanagem.
- O modelo só deve ser usado para realizar predições no intervalo de X ensaiado (de 1 a 6%, no exemplo), pois não temos informações sobre o relacionamento entre X e Y fora deste intervalo.
- O coeficiente a fornece uma estimativa do valor de Y quando X=0.

Equação da reta $\hat{y} = 79,7 + 0,886x$

Qualidade do ajuste

- Ajustou-se uma equação de regressão entre X e Y.
- Mas qual é a qualidade do ajuste?
  - Análise de variância do modelo
  - Análise dos resíduos

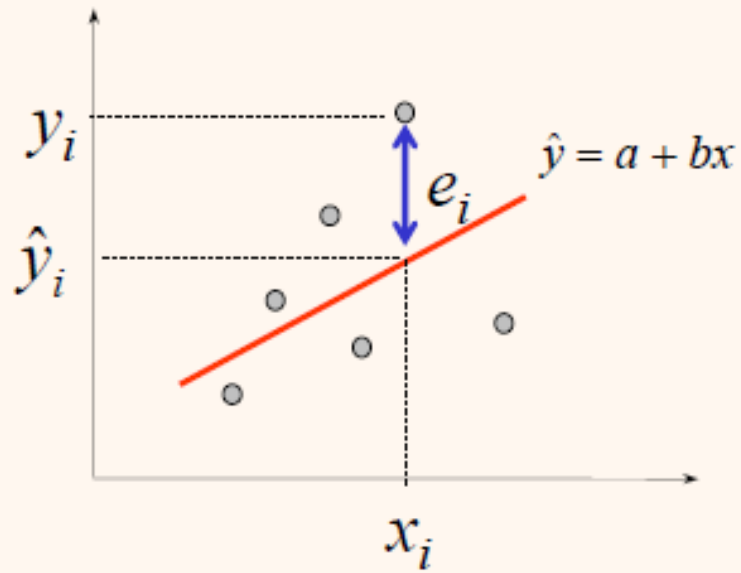
Reta de regressão e resíduos

- Valores preditos:

$$\hat{y}_i = a + bx_i$$

- Resíduos:

$$e_i = y_i - \hat{y}_i$$



Se X não influencia Y, então o valor esperado de Y pode ser estimado simplesmente pela média das observações de Y.

Mas se existe influência de X em Y, então deve haver algum ganho em considerar a equação da reta. Este ganho pode ser avaliado ao comparar os resíduos nas duas situações.

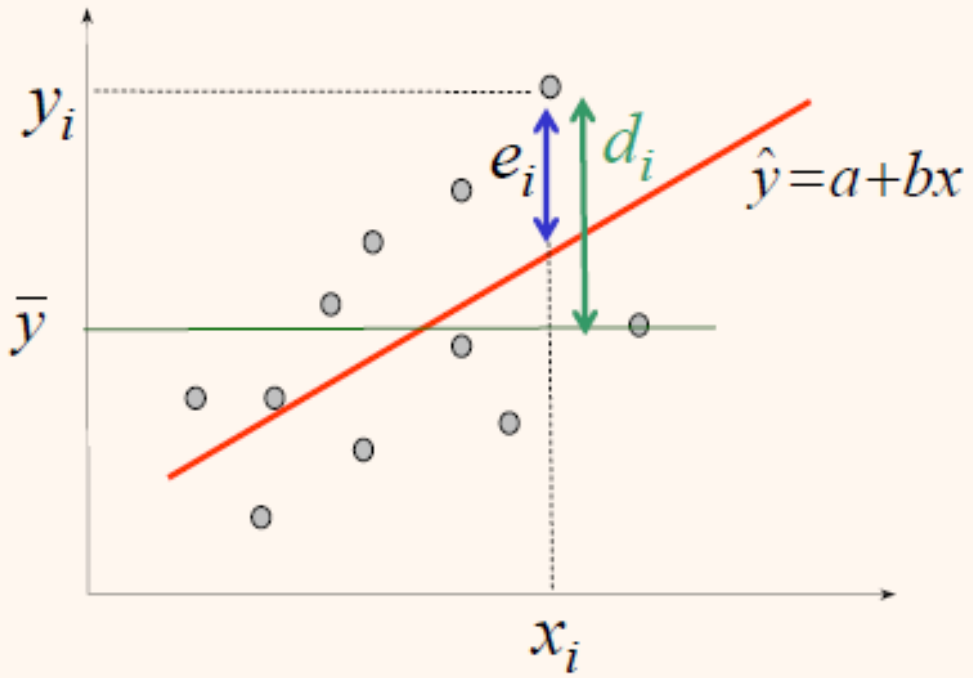
Análise de variância do modelo

Desvio em relação à média aritmética:

$$d_i = y_i - \bar{y}$$

Desvio em relação à reta de regressão (resíduo da regressão):

$$e_i = y_i - \hat{y}_i$$



As somas dos quadrados satisfazem a seguinte equação:

Somas de quadrados

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

SQT	SQR	SQE
variação total	variação explicada pela equação de regressão	variação não explicada

Processo simplificado de cálculo:

Somas de quadrados

$$SQT = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$SQE = \sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - a\sum y_i - b\sum x_i\, y_i$$

$$SQR = SQT - SQE$$

Coeficiente de determinação: $$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

O coeficiente de determinação é uma medida descritiva da proporção da variação de Y que pode ser explicada pela equação da reta

Medida da qualidade do ajuste:

Coeficiente de determinação (R^2)

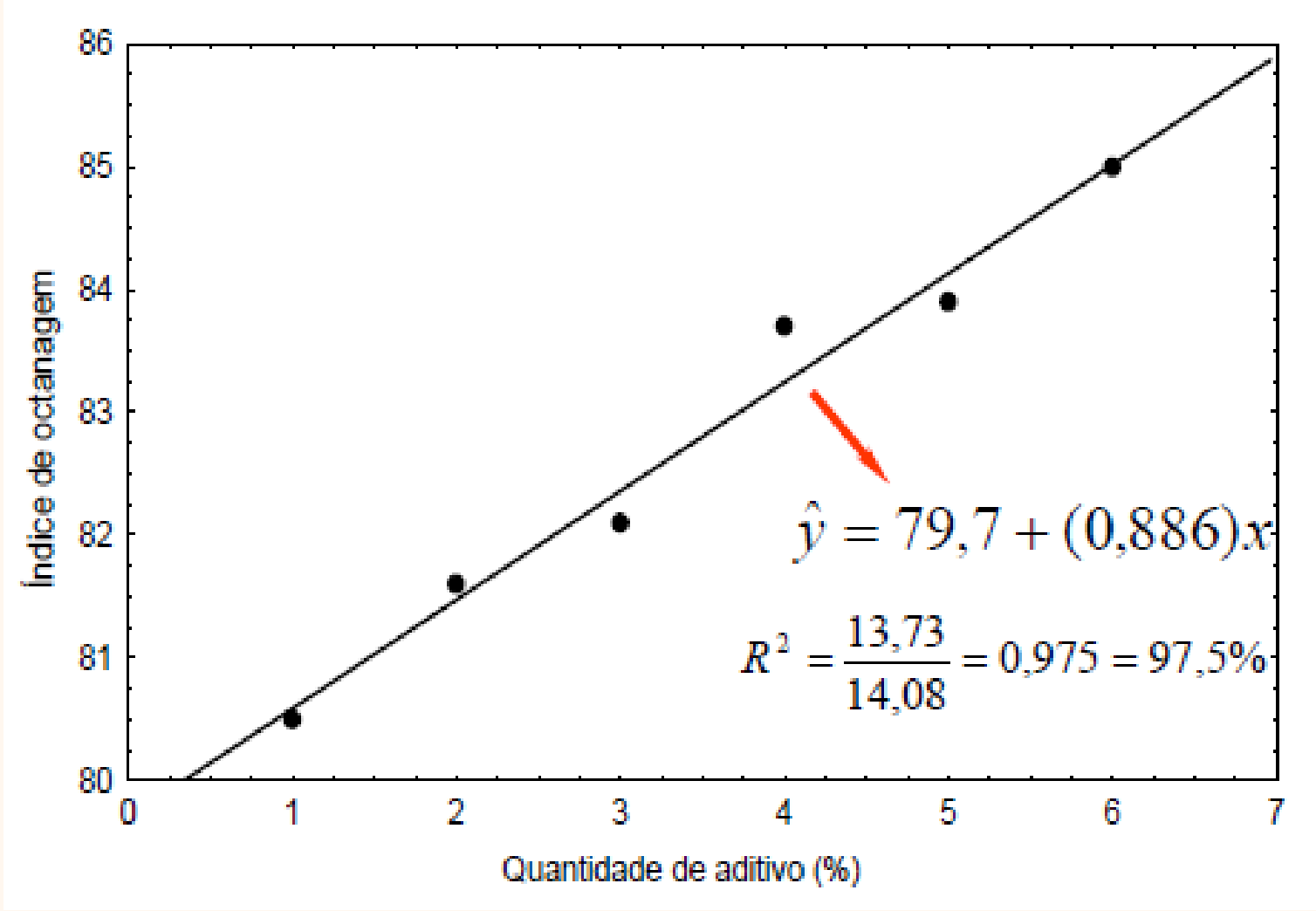
$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\Sigma (\hat{y}_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2}$$

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

Matematicamente, R^2 é o quadrado do Coef. de Correlação de Pearson.

No exemplo da octanagem da gasolina e a % de aditivos, $R^2 = 97,5\%$.

Interpretação: a variância da octanagem da gasolina é explicada em parte pela variação da quantidade de aditivo adicionada (97,5%) e em parte (2,5%) devido a outros fatores envolvidos no processo.



Os desvios de cada observação em relação às estimativas de E{Y} tem graus de liberdade iguais à n subtraído do número de parâmetros estimados em E{Y}.

Assim,
os desvios y_i–média tem n-1 graus de liberdade e
os desvios y_i–$y_{estimado}$ têm n-2 graus de liberdade.

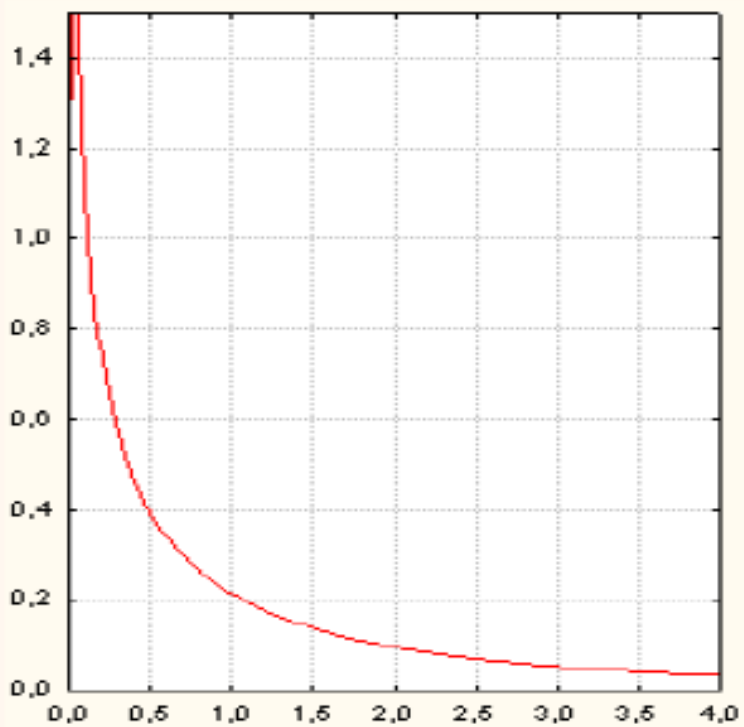
Análise de variância do modelo

Fonte de variação	*gl*	*SQ*	*QM*	Razão *f*
Regressão	1	$SQR = \sum(\hat{y}_i - \bar{y})^2$	$QMR = SQR/1$	$f = QMR/QME$
Erro	$n-2$	$SQE = \sum(y_i - \hat{y}_i)^2$	$QME = SQE/n-2$	
Total	$n-1$	$SQT = \sum(y_i - \bar{y})^2$		

Fonte de variação	*gl*	*SQ*	*MQ*	Razão *f*
Regressão	1	13,73	13,729	156,26
Erro	4	0,35	0,088	
Total	5	14,08		

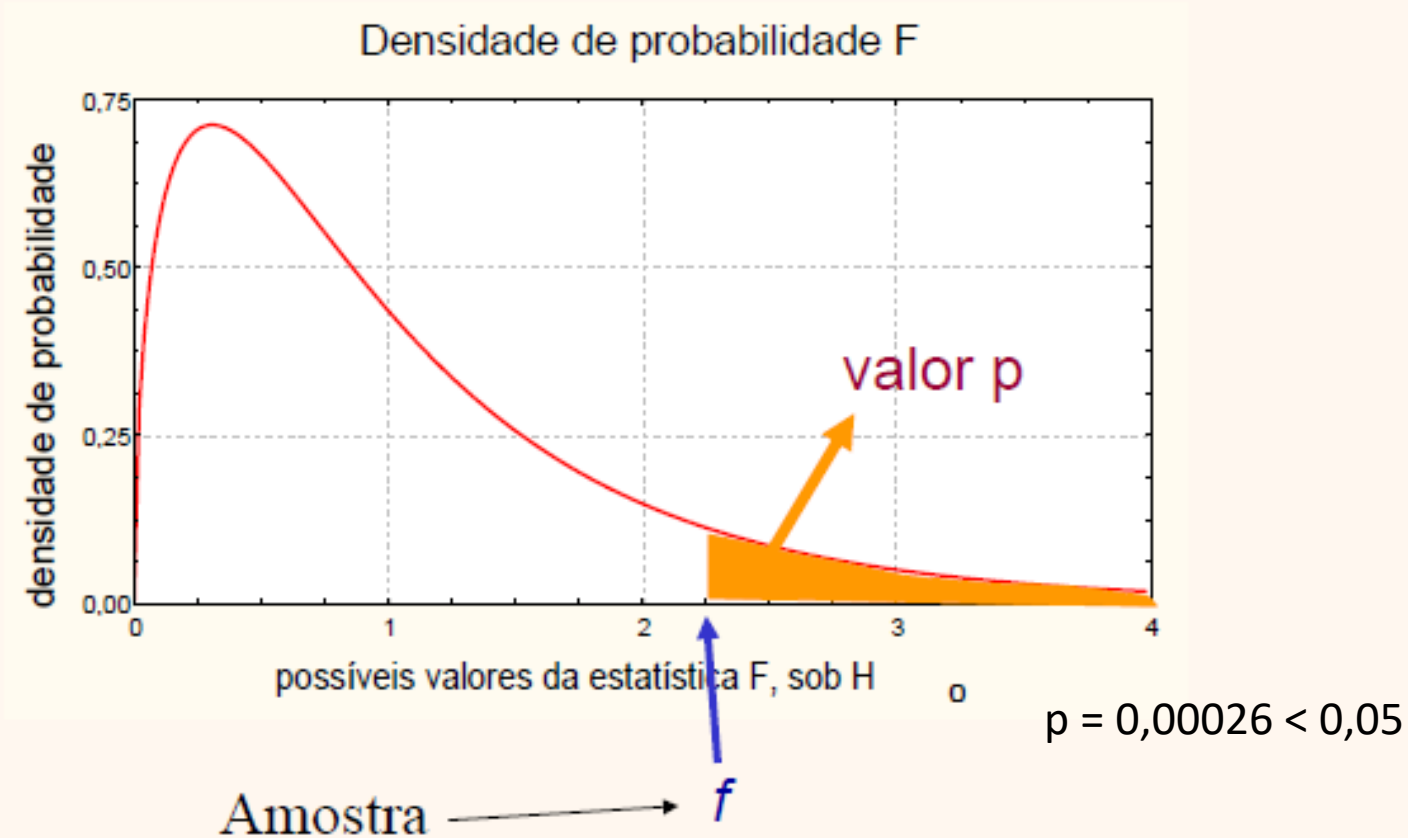
Usar a Tabela 6 e fazer o teste de significância do modelo.

Distribuição f com gl = 1 e 4



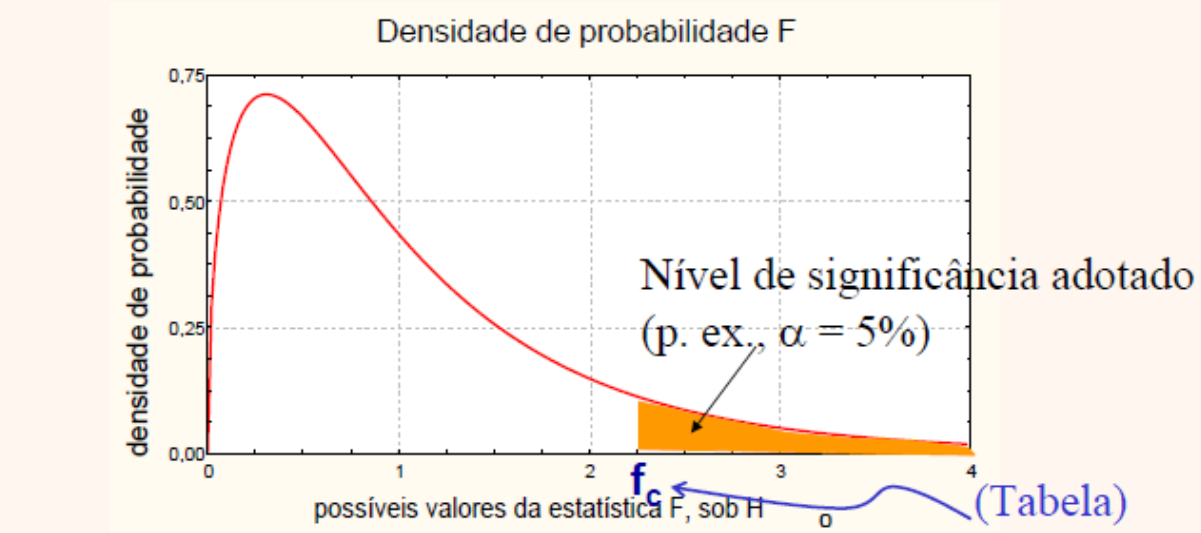
Possíveis valores de f, sob H_0

Valor p na distribuição F



p = 0,00026 < 0,05

Abordagem clássica: regra de decisão



$f_{\text{alfa=5\%; gl num =1, gl denom = 4}} = 7{,}7$

Logo rejeito H_0, ou seja, o modelo da reta é melhor que simplesmente a média.

Análise dos Resíduos

- Resíduo é a diferença R = Y - $\hat{Y}$
- Para verificar a adequação do ajuste deve-se construir o gráfico dos resíduos padronizados : $\frac{R}{s_R}$
- Se os pontos estiverem distribuídos dentro do intervalo [-2;+2], é uma indicação que o modelo está bem ajustado

Resíduo

Exemplo – índice de octanagem (Y) e quantidade de aditivos (X):

- Não há nenhuma observação fora do intervalo [-2;+2]:
- Espera-se menos de 5% fora do intervalo

aditivo (%)	Índice octanagem	octanagem estimada	residuo	resíduo padronizado
1	80,5	80,59	0,09	0,32
2	81,6	81,47	-0,13	-0,49
3	82,1	82,36	0,26	0,96
4	83,7	83,24	-0,46	-1,74
5	83,9	84,13	0,22	0,85
6	85	85,01	0,01	0,04
			0,27	

Exemplo 2: criminalidade e analfabetismo

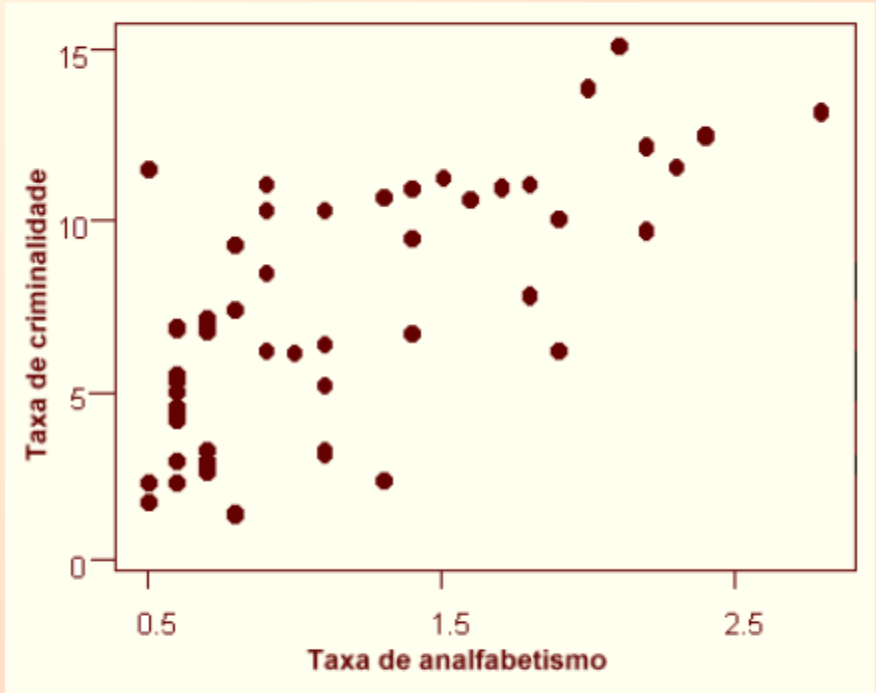
Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: taxa de criminalidade

X: taxa de analfabetismo

Obs.	Estado	Tanalf-70	Exvida-70	Tcrime-75	Obs.	Estado	Tanalf-70	Exvida-70	Tcrime-75
1	Alabama	2.1	69.05	15.1	26	Montana	0.6	70.56	5
2	Alaska	1.5	69.31	11.3	27	Nebraska	0.6	72.6	2.9
3	Arizona	1.8	70.55	7.8	28	Nevada	0.5	69.03	11.5
4	Arkansas	1.9	70.66	10.1	29	New-Hampshire	0.7	71.23	3.3
5	California	1.1	71.71	10.3	30	New-Jersey	1.1	70.93	5.2
6	Colorado	0.7	72.06	6.8	31	New-Mexico	2.2	70.32	9.7
7	Connecticut	1.1	72.48	3.1	32	New-York	1.4	70.55	10.9
8	Delaware	0.9	70.06	6.2	33	North-Carolina	1.8	69.21	11.1
9	Florida	1.3	70.66	10.7	34	North-Dakota	0.8	72.78	1.4
10	Georgia	2	68.54	13.9	35	Ohio	0.8	70.82	7.4
11	Hawaii	1.9	73.6	6.2	36	Oklahoma	1.1	71.42	6.4
12	Idaho	0.6	71.87	5.3	37	Oregon	0.6	72.13	4.2
13	Illinois	0.9	70.14	10.3	38	Pennsylvania	1	70.43	6.1
14	Indiana	0.7	70.88	7.1	39	Rhode-Island	1.3	71.9	2.4
15	Iowa	0.5	72.56	2.3	40	South-Carolina	2.3	67.96	11.6
16	Kansas	0.6	72.58	4.5	41	South-Dakota	0.5	72.08	1.7
17	Kentucky	1.6	70.1	10.6	42	Tennessee	1.7	70.11	11
18	Louisiana	2.8	68.76	13.2	43	Texas	2.2	70.9	12.2
19	Maine	0.7	70.39	2.7	44	Utah	0.6	72.9	4.5
20	Maryland	0.9	70.22	8.5	45	Vermont	0.6	71.64	5.5
21	Massachusetts	1.1	71.83	3.3	46	Virginia	1.4	70.08	9.5
22	Michigan	0.9	70.63	11.1	47	Washington	0.6	71.72	4.3
23	Minnesota	0.6	72.96	2.3	48	West-Virginia	1.4	69.48	6.7
24	Mississippi	2.4	68.09	12.5	49	Wisconsin	0.7	72.48	3
25	Missouri	0.8	70.69	9.3	50	Wyoming	0.6	70.29	6.9

Diagrama de dispersão



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a taxa de criminalidade (Y) tende a aumentar. Nota-se também uma tendência linear.

No exemplo 2,

a reta ajustada é:

$$\hat{Y} = 2{,}397 + 4{,}257\,X$$

$\hat{Y}$: valor predito para a taxa de criminalidade

X : taxa de analfabetismo

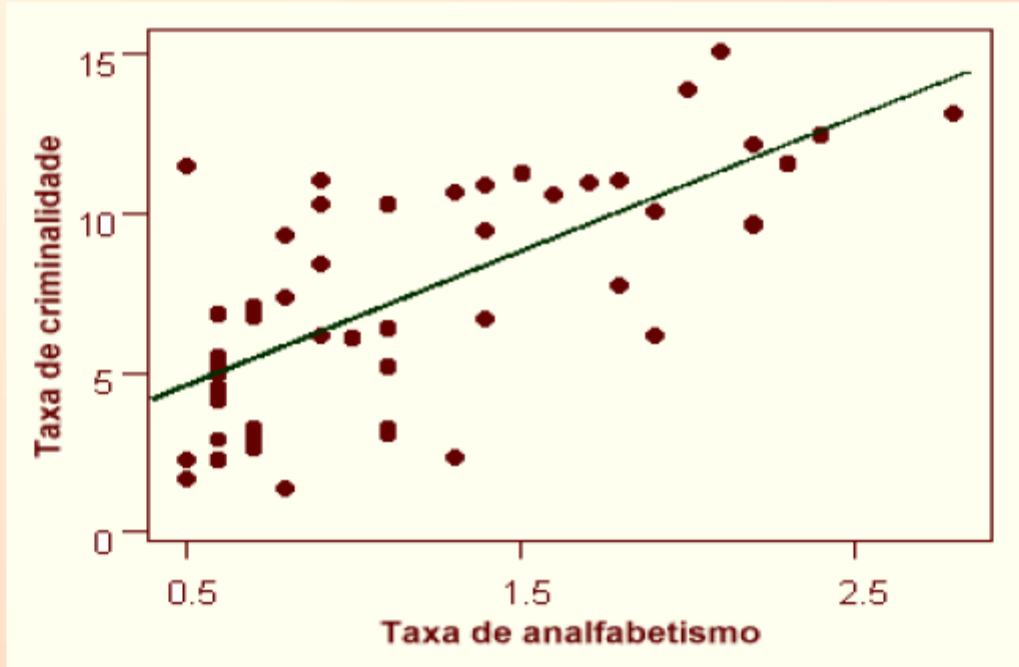
Interpretação de b:

Para um aumento de uma unidade na taxa do analfabetismo (X), a taxa de criminalidade (Y) aumenta, em média, 4,257 unidades.

Coeficiente de correlação de Pearson (r) = 0,702
Coeficiente de determinação (r^2) = 0,49 = 49%

Logo : 49% da variação da taxa de criminalidade deve-se à variação da taxa de analfabetismo

Graficamente, temos



Como desenhar a reta no gráfico?

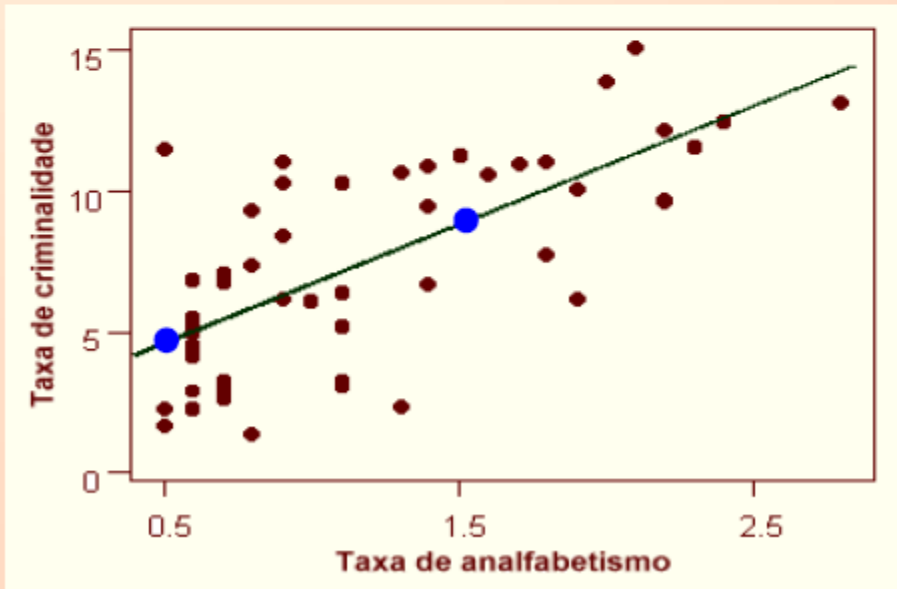
Escolha dois pontos:

- **X=0,5:**

$$\hat{y} = 2{,}397 + 4{,}257 \times 0{,}5 = 4{,}5255 \Rightarrow (0{,}5;4{,}53)$$

- **X=1,5:**

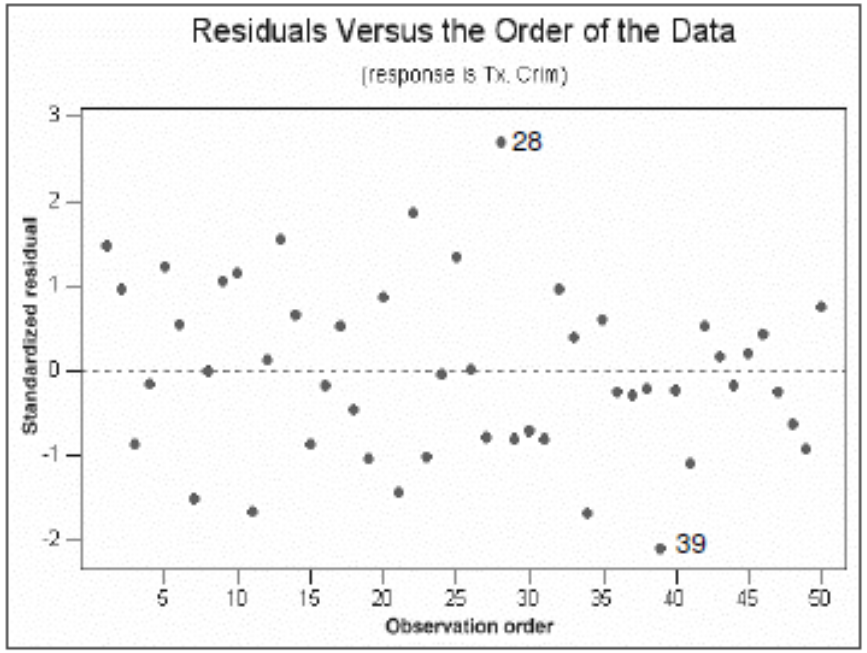
$$\hat{y} = 2{,}397 + 4{,}257 \times 1{,}5 = 8{,}7825 \Rightarrow (1{,}5;8{,}78)$$



Resíduo

Exemplo 2 - taxa de criminalidade (Y) e taxa de analfabetismo (X):

- Há 2 observações fora do intervalo [-2;+2]:
  - #28= Nevada
  - #39= Rhode-Island
- São considerados valores aberrantes
- Espera-se menos de 5% fora do intervalo



Resíduo

Exemplo 2 - taxa de criminalidade (Y) e taxa de analfabetismo (X), eliminando-se o Estado de Nevada:

- Correlação com todos os estados r = 0,702
- Correlação sem Nevada r = 0,748

- Equação com todos os estados: $\hat{Y}$ = 2,397 + 4,257X
- Equação sem Nevada: $\hat{Y}$ = 1,936 + 4,526X

Voltando ao exemplo 3

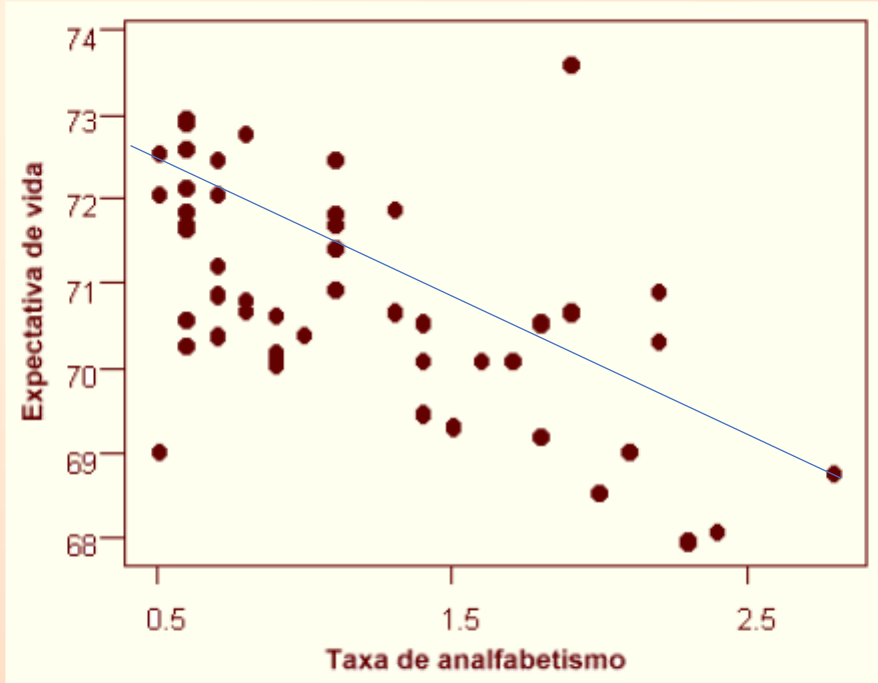
Exemplo 3: expectativa de vida e analfabetismo

Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: expectativa de vida

X: taxa de analfabetismo

Diagrama de dispersão



Podemos notar que, conforme aumenta a taxa de analfabetismo (X), a expectativa de vida (Y) tende a diminuir. Nota-se também uma tendência linear.

No exemplo 3,

a reta ajustada é:

$$\hat{Y} = 72{,}395 - 1{,}296\ X$$

$\hat{Y}$: valor predito para a expectativa de vida

X : taxa de analfabetismo

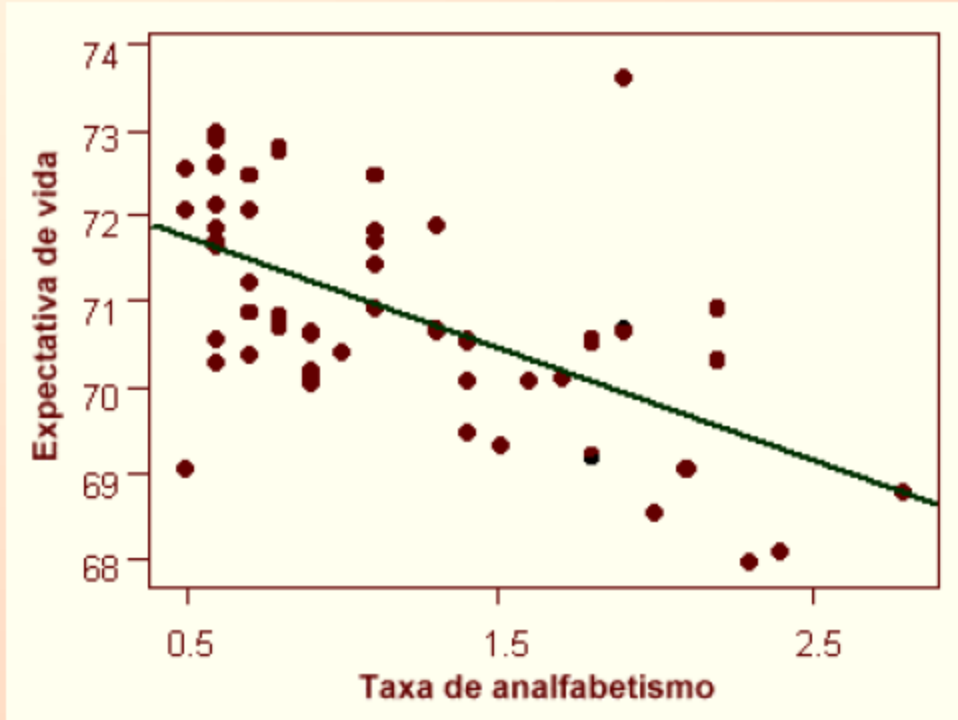
Interpretação de b:

Para um aumento de uma unidade na taxa do analfabetismo (X), a expectativa de vida (Y) diminui, em média, 1,296 anos.

Coeficiente de correlação de Pearson (r) = - 0,59
Coeficiente de determinação (r^2) = 0,35 = 35%

Logo : 35% da variação da expectativa de vida deve-se à variação da taxa de analfabetismo

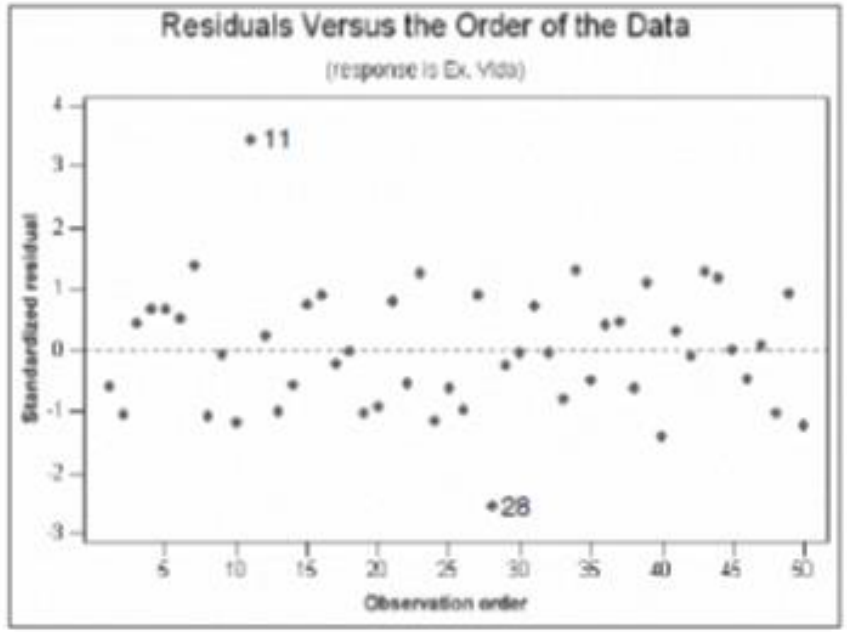
Graficamente, temos



Resíduo

Exemplo 3 - expectativa de vida (Y) e taxa de analfabetismo (X):

- Há 2 observações fora do intervalo [-2;+2]:
  - #28= Nevada
  - #11= Hawaii
- São considerados valores aberrantes
- Espera-se menos de 5% fora do intervalo



Resíduo

Exemplo 3 - Expectativa de vida (Y) e Taxa de analfabetismo (X), eliminando-se os Estados de Nevada e Hawaii:

- Correlação com todos os estados r = -0,590
- Correlação sem Nevada e sem Hawaii r = -0,797

- Equação da reta com todos os estados: $\hat{Y}$ = 72,395 – 1,296X
- Equação sem Nevada e sem Hawaii : $\hat{Y}$ = 72,680 – 1,557X

obrigada