Universidade Federal de Mato Grosso

AULA 08 – Estatística

Prof. Lucas Bianchi

Cuiabá, 12 de setembro de 2016

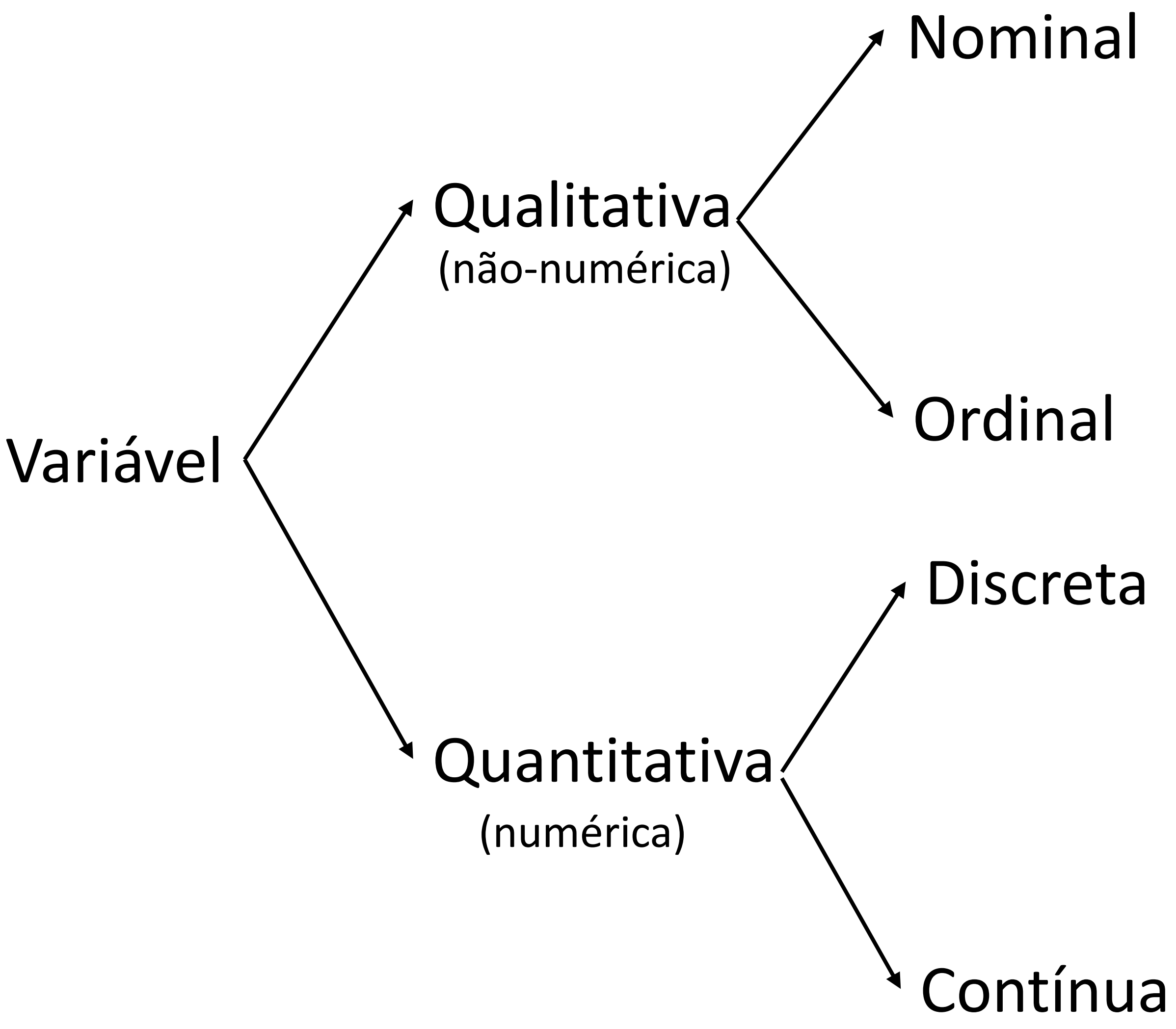
Na aula de hoje, estudaremos:

- ✓ Variáveis aleatórias
- ✓ Correlação
- ✓ Regressão Linear

Introdução

Existem situações nas quais há interesse em estudar o comportamento conjunto de uma ou mais variáveis. Em muitos casos, a explicação de um fenômeno de interesse pode estar associado a outros fatores (variáveis) que contribuem de algum modo para a ocorrência deste fenômeno.

Classificação de variáveis



Qualitativa Nominal

Consistem apenas em nomes, rótulos ou categorias. Os dados não podem ser dispostos segundo um esquema ordenado.

Ex: Masculino, Feminino.

Qualitativa Ordinal

Envolve dados que podem ser dispostos em alguma ordem, mas as diferenças entre os valores dos dados não podem ser determinadas ou não tem sentido.

Ex: Ens. Fudamental, médio e superior.

Quantitativa Discreta

Assume valores pertencentes a um conjunto finito ou enumerável. Geralmente, seus valores são resultados de um processo de contagem, razão pela qual seus valores são expressos através de números inteiros não-negativos.

Ex: Quantidade de membros por familia.

Quantitativa Contínua

Assume qualquer valor pertencente a um determinado intervalo do conjunto dos Reais. Pode-se dizer que a variável contínua resulta normalmente de mensurações.

Ex: Nota, Altura, Peso.

Associação entre variáveis Quantitativas

Quando as variáveis envolvidas são ambas do tipo quantitativo é possível utilizar procedimentos analíticos como:

- Gráfico de dispersão
- Correlação

Utilizamos o diagrama de dispersão entre duas variáveis para:

- Mostrar a relação entre duas variáveis quantitativas, medidas sobre os mesmos indivíduos;
- Os valores de uma variável aparecem no eixo horizontal, e os da outra, no eixo vertical.

Cada indivíduo aparece como o ponto do gráfico definido pelos valores de ambas as variáveis para aquele indivíduo

Alguns aspectos são importantes ao se analisar o gráfico de dispersão, são eles:

- Direção (crescente, decrescente)
- Forma (linear, não linear, aglomerados)
- Pontos discrepantes

Tabela 1: Anos de serviço (X) por Números de clientes (Y)

Agente	Anos de serviço (X)	Número de clientes (Y)
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

Figura 1: Gráfico de dispersão para as variáveis anos de serviço e número de clientes

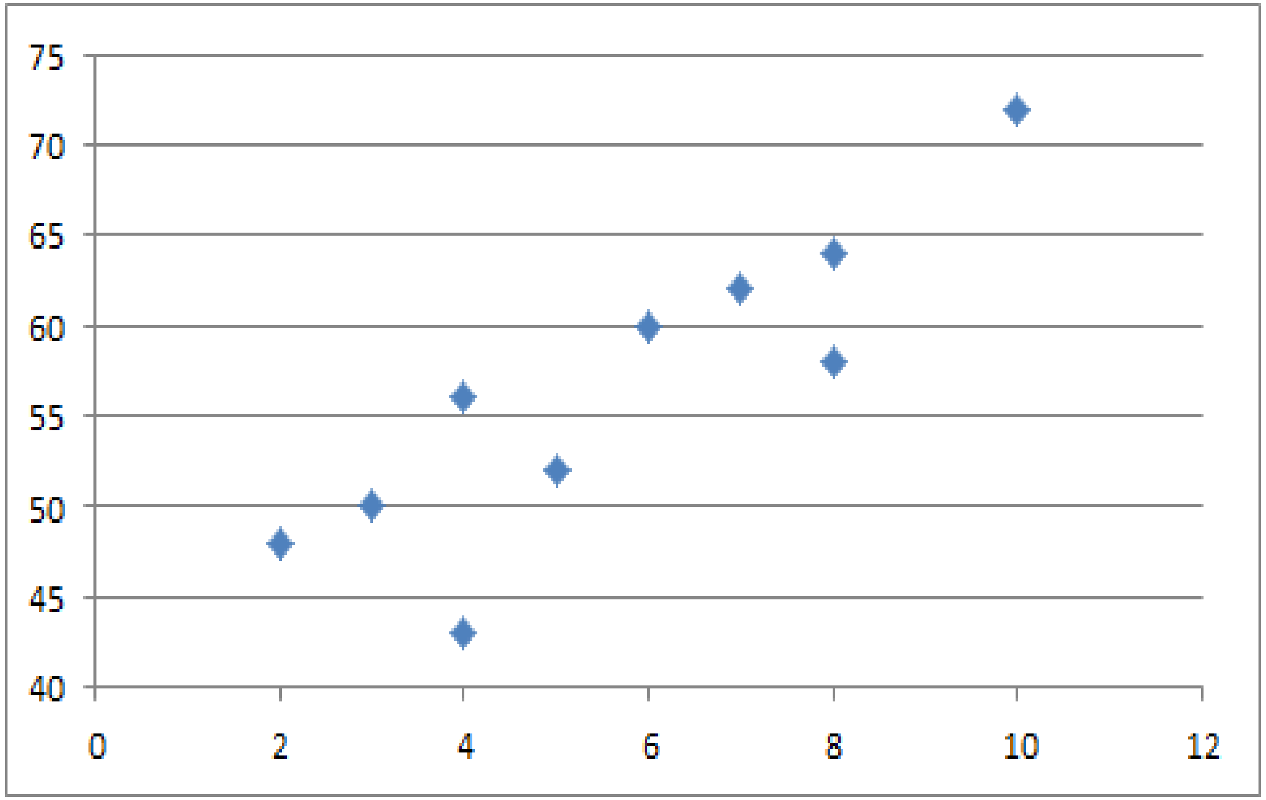


Tabela 2: Anos de serviço (X) por Números de clientes (Y)

Variável	X	Y
A	3	44
B	10	19
C	4	40
D	9	42
E	1	17
F	4	20
G	7	26
H	10	20
I	4	25
J	4	31

Figura 2: Gráfico de dispersão para as variáveis anos de serviço e número de clientes

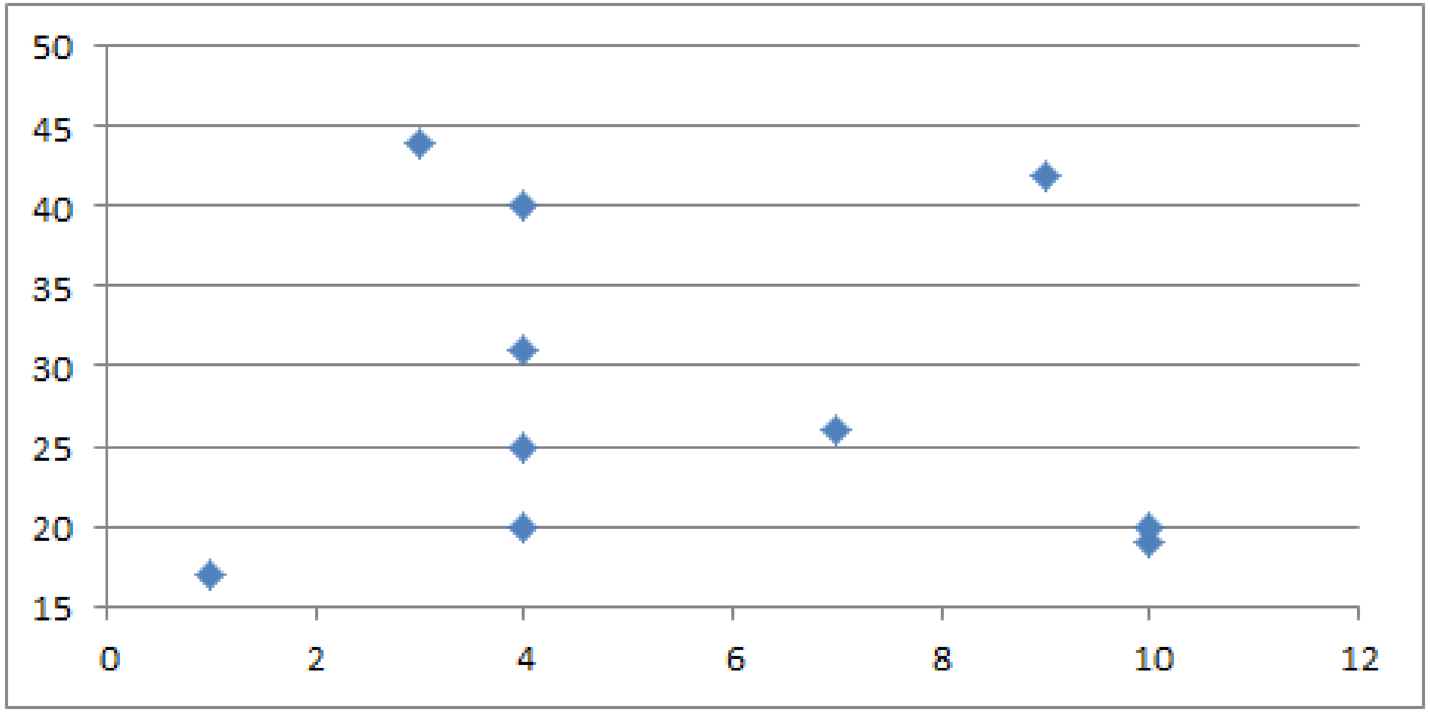
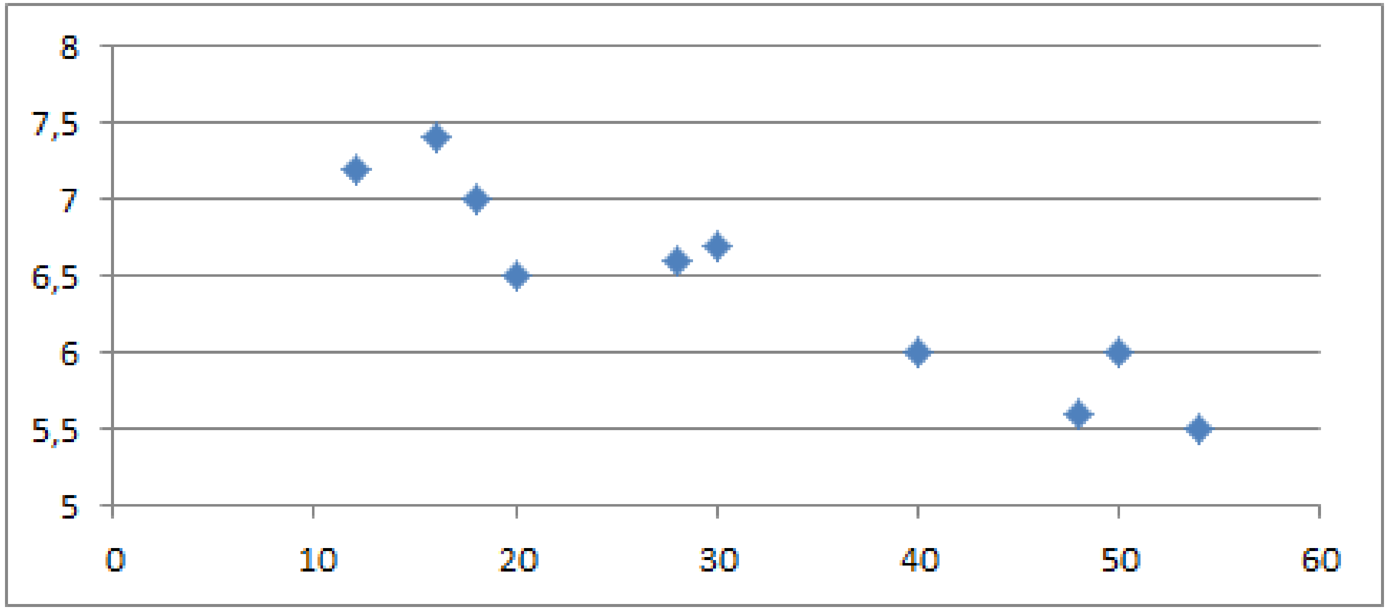


Tabela 3: Renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y)

Familia	X	Y
A	12	7,2
B	16	7,4
C	18	7
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6
H	48	5,6
I	50	6
J	54	5,5

Figura 3: Gráfico de dispersão para as variáveis X e Y



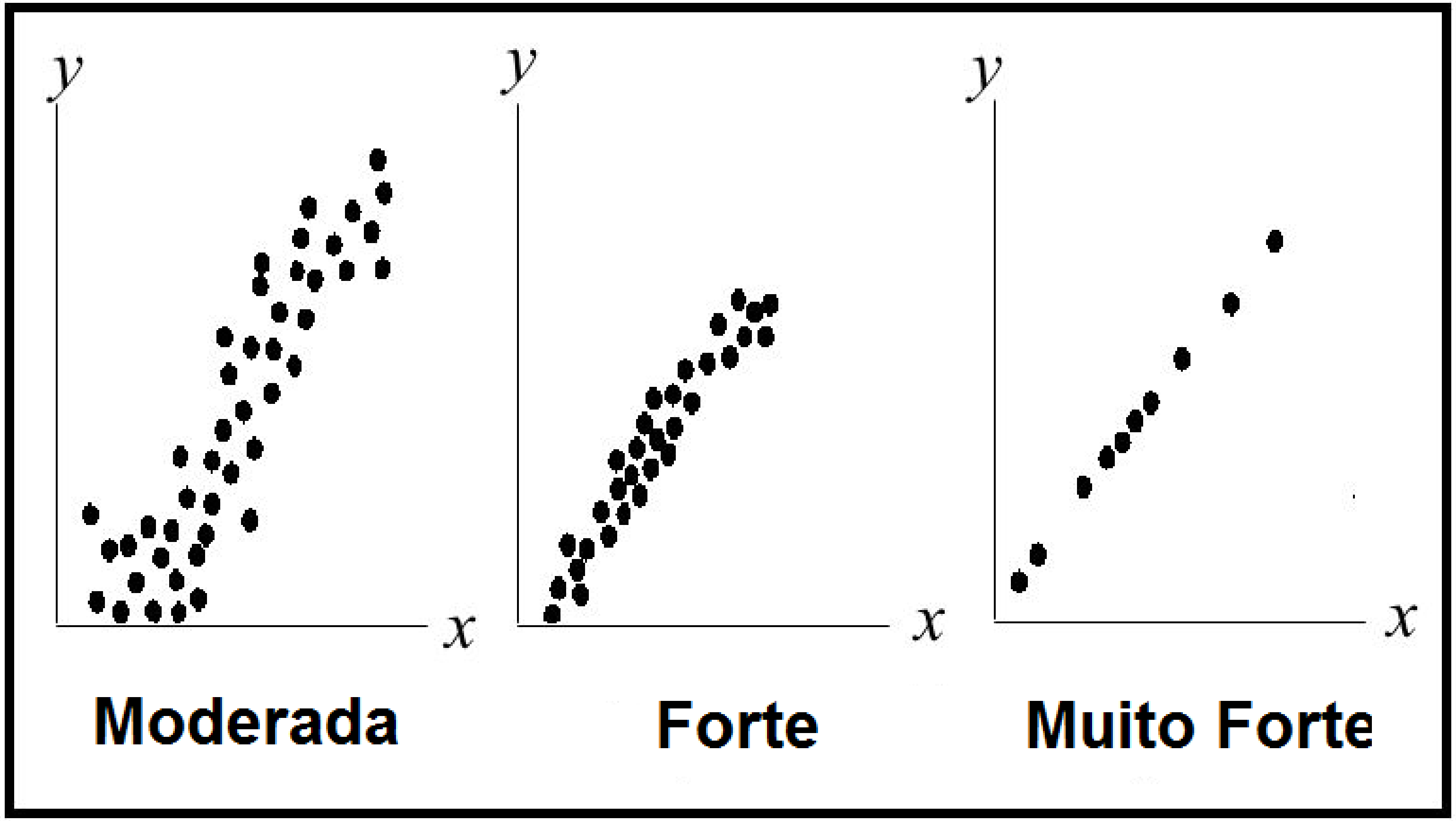
Correlação Linear Simples

Vimos que os gráficos ajudam a verificar, de forma rápida e fácil, a existência de possíveis associações. Entretanto, é muito difícil quantificar essa relação entre a variável X e Y apenas olhando o gráfico. Portanto, precisamos definir uma medida que quantifica essa associação.

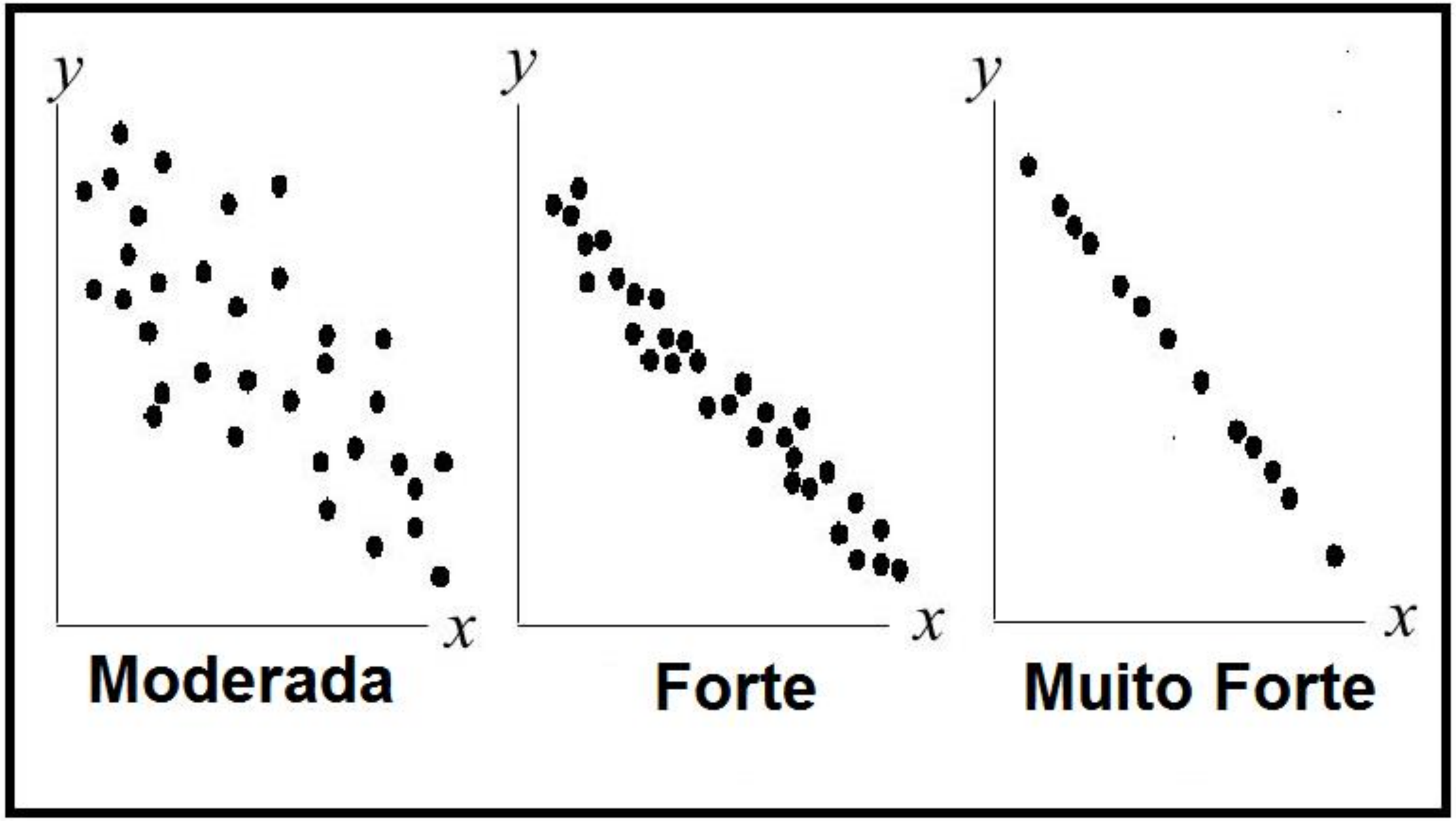
A relação mais simples que podemos observar entre duas variáveis é a relação linear. Assim, teremos uma medida que quantifica o quanto os pontos se aproximam de uma reta.

Essa medida será definida de modo a variar num intervalo finito, especificamente, de -1 a +1.

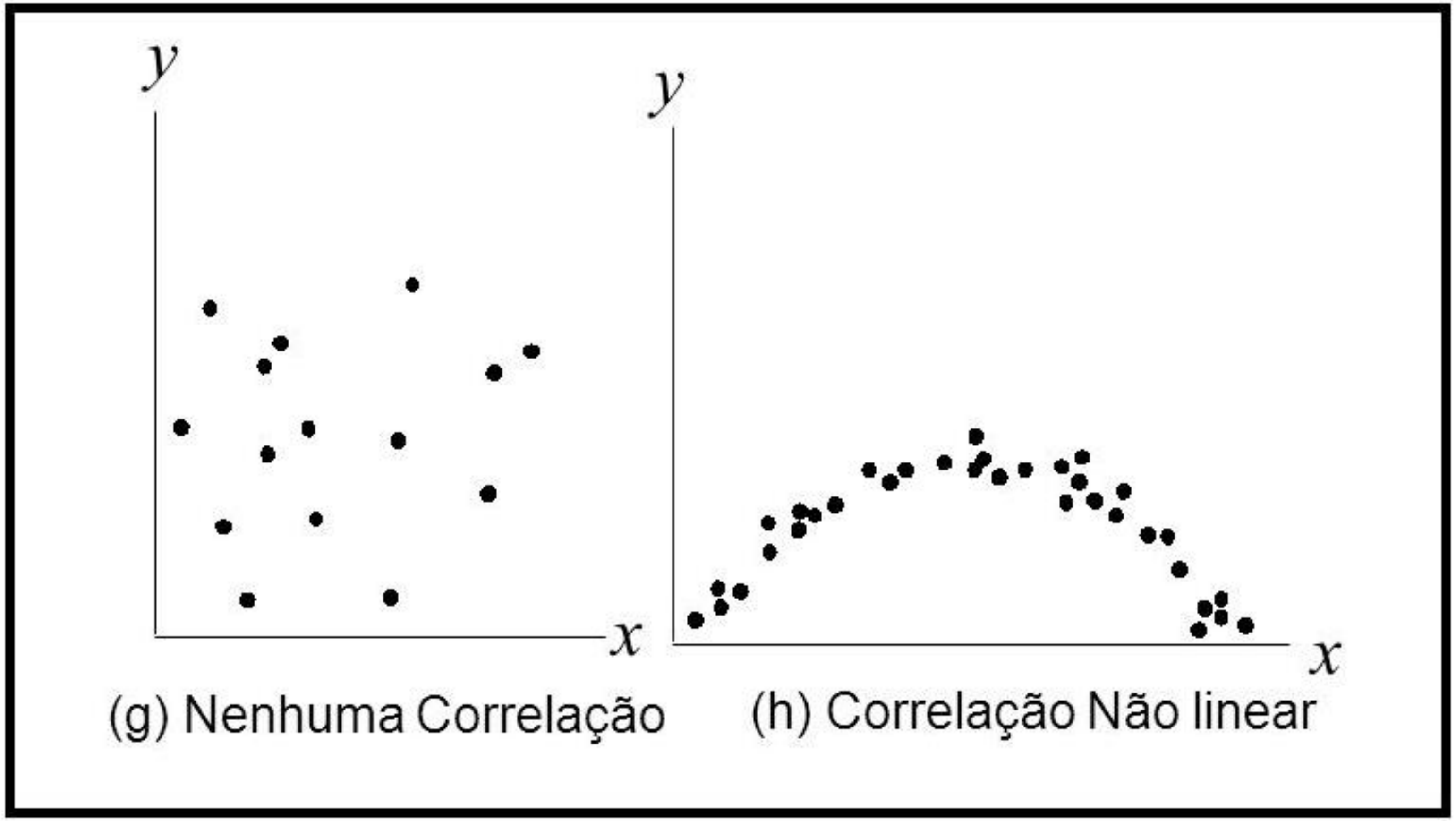
Correlação Positiva Linear



Correlação Negativa Linear



Correlação Não Linear



Utilizamos a correlação linear entre duas variáveis para:

- Verificar se uma delas está, de alguma forma, relacionada com a outra;
- Analisar se a alteração no valor de uma variável (dita independente) provoca alterações no valor da outra variável (dita dependente)

Para fazer isso precisamos calcular

Tabela 4: Anos de serviço (X) por Números de clientes (Y)

Agente	Anos (X)	Clientes (Y)	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})\,(y - \bar{y})$	$(x - \bar{x})^{2}$	$(y - \bar{y})^{2}$
A	2	48	-3,70	-8,50	31,45	13,69	72,25
B	3	50	-2,70	-6,50	17,55	7,29	42,25
C	4	56	-1,70	-0,50	0,85	2,89	0,25
D	5	52	-0,70	-4,50	3,15	0,49	20,25
E	4	43	-1,70	-13,50	22,95	2,89	182,25
F	6	60	0,30	3,50	1,05	0,09	12,25
G	7	62	1,30	5,50	7,15	1,69	30,25
H	8	58	2,30	1,50	3,45	5,29	2,25
I	8	64	2,30	7,50	17,25	5,29	56,25
J	10	72	4,30	15,50	66,65	18,49	240,25
Total	57	565			171,5	58,1	658,5

Até agora, fizemos cálculos necessários para obter o coeficiente de correlação entre as duas variáveis X e Y. A formula desse cálculo é dada por:

$$corr(X,Y) = \frac{\boxed{\sum(x_i - \bar{x})(y_i - \bar{y})}}{\sqrt{\boxed{\sum(x_i - \bar{x})^2}\boxed{\sum(y_i - \bar{y})^2}}} = \frac{171,5}{\sqrt{58,1 * 658,5}} = 0,8767$$

Observa-se que temos todos os campos necessários para obtermos o valor da correlação, desta forma, basta fazer as substituições

Portanto, neste exemplo, o grau de correlação linear é de 87,67%, ou seja, há uma correlação positiva forte.

Tabela 5: Renda bruta mensal (X) e porcentagem da renda gasta em saúde (Y)

Agente	Anos (X)	Clientes (Y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})\,(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
A	12	7,2	-19,6	0,75	-14,7	384,16	0,5625
B	16	7,4	-15,6	0,95	-14,82	243,36	0,9025
C	18	7	-13,6	0,55	-7,48	184,96	0,3025
D	20	6,5	-11,6	0,05	-0,58	134,56	0,0025
E	28	6,6	-3,6	0,15	-0,54	12,96	0,0225
F	30	6,7	-1,6	0,25	-0,4	2,56	0,0625
G	40	6	8,4	-0,45	-3,78	70,56	0,2025
H	48	5,6	16,4	-0,85	-13,94	268,96	0,7225
I	50	6	18,4	-0,45	-8,28	338,56	0,2025
J	54	5,5	22,4	-0,95	-21,28	501,76	0,9025
Total	316	64,5			$\boxed{-85,8}$	$\boxed{2142,4}$	$\boxed{3,885}$

Uma vez feito os cálculos, vamos obter o grau de associação linear (correlação)

$$corr(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{-85,5}{\sqrt{2142,4 * 3,885}} = -0,9404$$

Observa-se que temos todos os campos necessários para obtermos o valor da correlação, desta forma, basta fazer as substituições

Portanto, neste exemplo, o grau de correlação linear é de -94,04%, ou seja, há uma correlação linear negativa forte.

Atenção!!

As propriedades mais importantes do coeficiente de correlação são:

1. O coeficiente de correlação é uma medida adimensional, isto é, ele é independente das unidades de medida das variáveis X e Y.
2. Quanto mais próximo de +1 for “r”, maior o grau de relacionamento linear positivo entre X e Y, ou seja, se X varia em uma direção, Y variará na mesma direção.
3. Quanto mais próximo de -1 for “r, maior o grau de relacionamento linear negativo entre X e Y, isto é, se X varia em um mesmo sentido Y, variará no sentido inverso.
4. Quanto mais próximo de zero estiver “r”, menor será o relacionamento linear entre X e Y.

Observação: R^2, chamado de coeficiente de determinação é dado simplesmente pelo quadrado de R.

Propriedades:

- Se R > 0, então a relação linear é positiva.
- Se R< 0, então a relação linear é negativa.
- Se R^2 é próximo de 1, então dizemos que a relação linear é forte.
- Se R^2 é próximo de 0, então dizemos que a relação linear é fraca.
- Se R^2 for intermediário, a relação linear é intermediária.

Regressão Linear

Descreve a relação entre duas variáveis por meio de uma reta. Essa reta é chamada de reta de regressão e é dada por:

$$y = a + xb + e_i$$

Em que “a” (é o intercepto) e “b" (inclinação da reta) são parâmetros do modelo, ou seja, para encontrarmos a reta de regressão. Podemos utilizar o método de mínimos quadrados para calcular essas constantes.

Mínimos Quadrados

É uma técnica de otimização matemática que procura o melhor ajuste para um conjunto de dados, ou seja, procura-se encontrar a reta que mais se aproxima dos dados.

O objetivo deste método é encontrar os valores para "a" e "b" que minimizem a soma do quadrado dos resíduos. Os resíduos são as diferenças entre o valor estimado e o observado.

Podemos calcular esses parâmetros utilizando as expressões abaixo:

$$a = \bar{y} - b\bar{x} \qquad \qquad \mathrm{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Exercício 1. (CAPES/2008) Considere as afirmações abaixo:

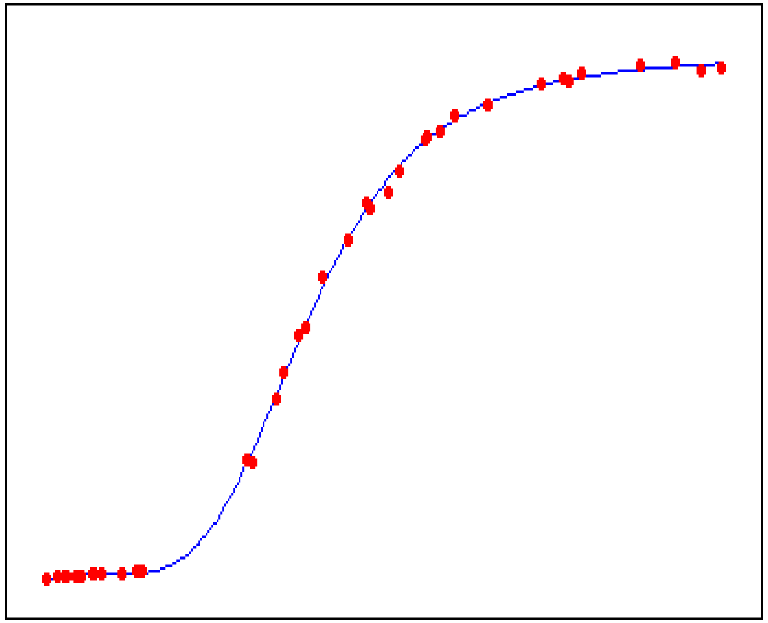
I) O coeficiente de correlação linear de Pearson é necessariamente um número no intervalo (-1,1).

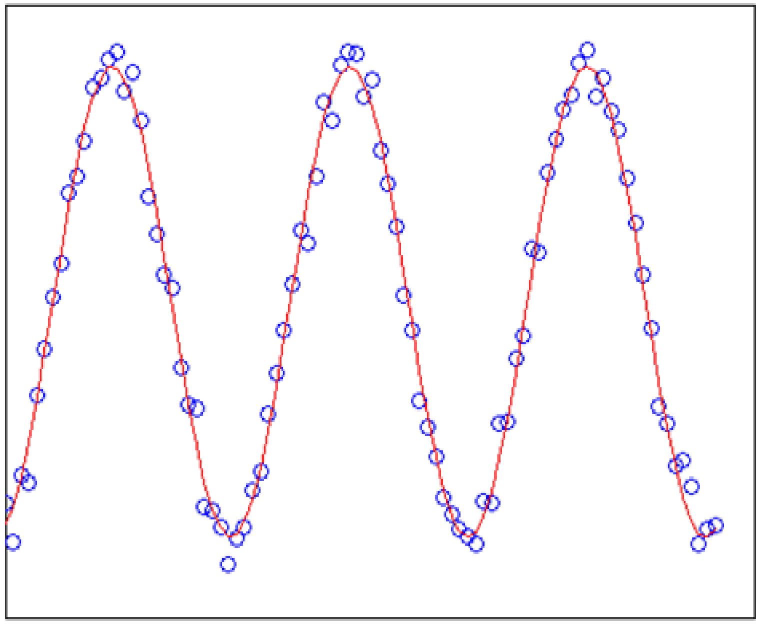
II) O coeficiente de correlação linear de Pearson só pode ser calculado para variáveis quantitativas.

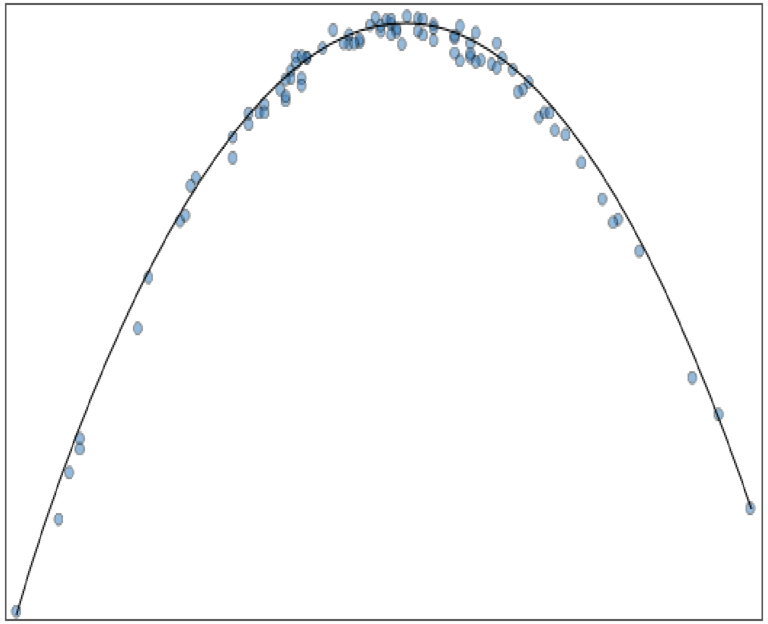
a) As duas afirmações são verdadeiras, e a segunda é uma justificativa correta da primeira
b) As duas são verdadeiras, e a segunda não é uma justificativa correta da primeira.
c) A primeira é verdadeira e a segunda é falsa.
d) A primeira é falsa e a segunda é verdadeira.
e) Ambas são falsas.

Regressão não-linear

A relação entre as variáveis não pode ser descrita por uma função linear como demonstrado nas figuras abaixo.

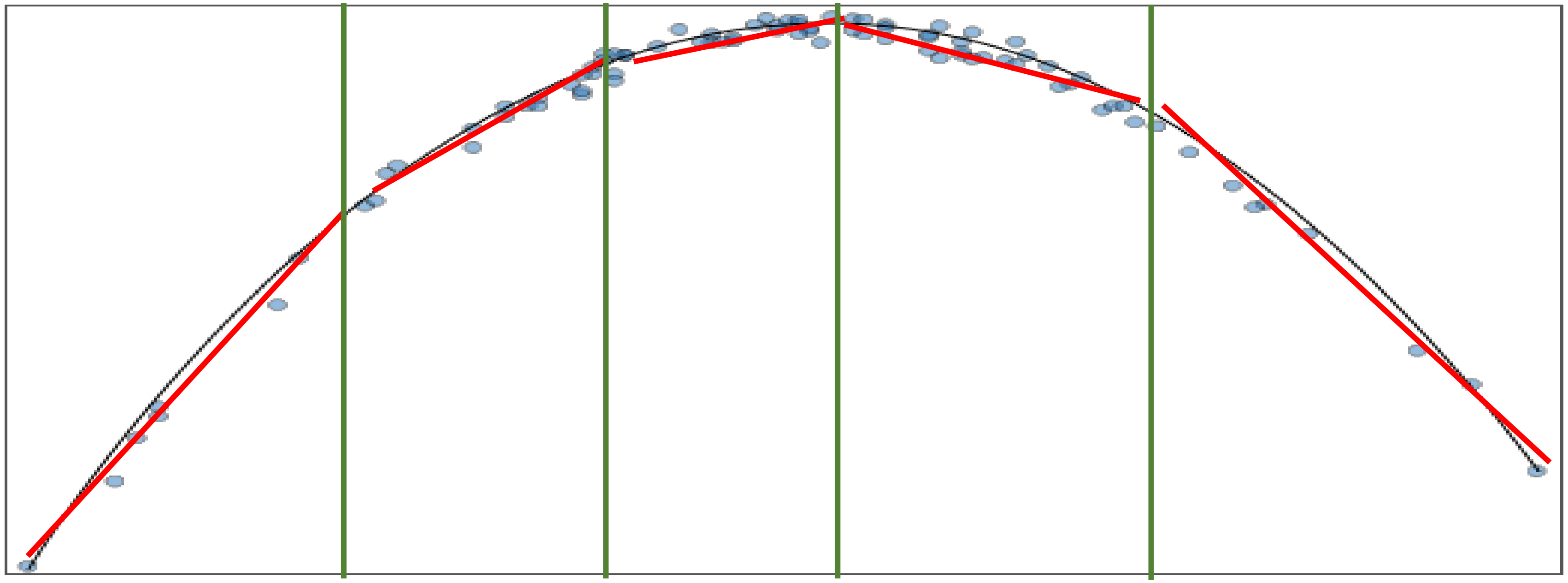






Mas o que fazer quando o modelo de regressão linear não é apropriado?

Solução 1: Dividir o dado em partes suficientes para que a regressão linear seja adequada, exemplo:



Mas o que fazer quando o modelo de regressão linear não é apropriado?

Solução 2: Podemos utilizar um modelo de regressão não linear.

Solução 3: Aplicar uma linearização, ou seja, transformar os dados para que seja possível continuar usando a regressão linear.

Regressão linear múltipla

Regressão múltipla é uma coleção de técnicas estatísticas para construir modelos que descrevem de maneira razoável relações entre várias variáveis explicativas de um determinado processo.

A diferença entre a regressão linear simples e a múltipla é que na múltipla são tratadas duas ou mais variáveis explicativas.

Exercício 2: Recordemos o exemplo em que se pretende estudar a relação entre o volume de vendas (Y) efetuadas durante um dado período de tempo por um vendedor, os seus anos de experiência (X1) e o seu score num teste de inteligência (X2)

Vendedor	Vendas (Y)	Anos de experiência (X1)	Score no teste de inteligência (X2)
1	9	6	3
2	6	5	2
3	4	3	2
4	3	1	1
5	3	4	1

Vendedor	Vendas (Y)	Anos de experiência (X1)	Score no teste de inteligência (X2)
6	5	3	3
7	8	6	3
8	2	2	1
9	7	4	2
10	4	2	2

Exercício 1. Um pesquisador deseja verificar se um instrumento para medir a concentração de determinada substância no sangue está bem calibrado. Para isto, ele tomou 15 amostras de concentrações conhecidas (X) e determinou a respectiva concentração através do instrumento (Y), obtendo:

X	2,0	2,0	2,0	4,0	4,0	4,0	6,0	6,0	6,0	8,0	8,0	8,0	10,0	10,0	10,0
Y	2,1	1,8	1,9	4,5	4,2	4,0	6,2	6,0	6,5	8,2	7,8	7,7	9,6	10,0	10,1

(a) Construa o diagrama de dispersão para esses dados.

(b) Calcule o coeficiente de correlação entre as variáveis X e Y.

(c) Obtenha a reta de regressão da variável Y em função de X.

Gabarito – Exercício 1

$R = 0{,}9961$

$R^2 = 0{,}9922$

X	Y	coluna 1	coluna 2	coluna 3	coluna 4	coluna 5
2	2,1	-4	-3,94	15,76	16	15,5236
2	1,8	-4	-4,24	16,96	16	17,9776
2	1,9	-4	-4,14	16,56	16	17,1396
4	4,5	-2	-1,54	3,08	4	2,3716
4	4,2	-2	-1,84	3,68	4	3,3856
4	4	-2	-2,04	4,08	4	4,1616
6	6,2	0	0,16	0	0	0,0256
6	6	0	-0,04	0	0	0,0016
6	6,5	0	0,46	0	0	0,2116
8	8,2	2	2,16	4,32	4	4,6656
8	7,8	2	1,76	3,52	4	3,0976
8	7,7	2	1,66	3,32	4	2,7556
10	9,6	4	3,56	14,24	16	12,6736
10	10	4	3,96	15,84	16	15,6816
10	10,1	4	4,06	16,24	16	16,4836
				117,6	120	116,156