Estatística e Probabilidade

Correlação e Regressão Linear

Desenvolvimento do material

Gregório Dalle Vedove Nosaki

1ª Edição

Sumário

Correlação e Regressão Linear

Para início de conversa...

Nosso último tópico sobre Estatística faz parte de uma das teorias mais utilizadas no estudo e na análise da relação entre duas variáveis. A partir do diagrama de dispersão, que é uma representação geométrica dos valores observados, podemos ter uma ideia inicial da existência de uma correlação entre as variáveis apresentadas. Veremos que tal correlação pode ser diretamente ou inversamente proporcional, e que tal característica é indicada pelo sinal do coeficiente de correlação linear de Pearson. Esse coeficiente, cujo módulo está sempre entre zero e um, indicará, ainda, o grau de correlação entre as duas variáveis, podendo ser classificado como baixo, moderado ou alto. Por fim, trabalharemos na construção e na determinação da equação da reta que melhor relaciona duas variáveis. Essa reta torna possível uma análise mais detalhada da relação entre os pontos que estamos considerando.

Trabalharemos, aqui, apenas com a correlação linear, mas existem outros tipos de relações entre duas variáveis, apresentando diversos exemplos para ilustrar as definições e os conceitos abordados.

Objetivo

Utilizar a correlação e a regressão linear para analisar a relação entre duas variáveis e realizar predições.

1. Diagrama de Dispersão

Neste capítulo, estaremos voltados para compreender a relação estabelecida entre duas variáveis. Trabalharemos com algumas das principais ferramentas utilizadas para estabelecer se há de fato uma relação entre elas e, até mesmo, estimar uma função que descreva tal relação. A partir de um conjunto de pares de observáveis, podemos construir um **diagrama de dispersão,** que nos dá uma visão geométrica de como os valores da nossa amostra estão dispostos em um plano cartesiano.

Considere os dados fictícios apresentados na tabela referente à idade e à altura de alguns indivíduos.

Nome	Idade (anos)	Altura (cm)
Alberto	10	138
Bruno	5	105
César	12	146
Diogo	8	122
Eduardo	9	131
Felipe	3	93
Gustavo	4	101
Heitor	10	128
Iago	9	135
Júlio	2	85

O primeiro passo para analisar se há uma correlação entre a idade e a altura dos indivíduos nessa amostra é a plotagem do diagrama de dispersão, em que os pares são ordenados formados pela idade e altura. Observe o diagrama de dispersão dos dados da tabela anterior apresentado a seguir.

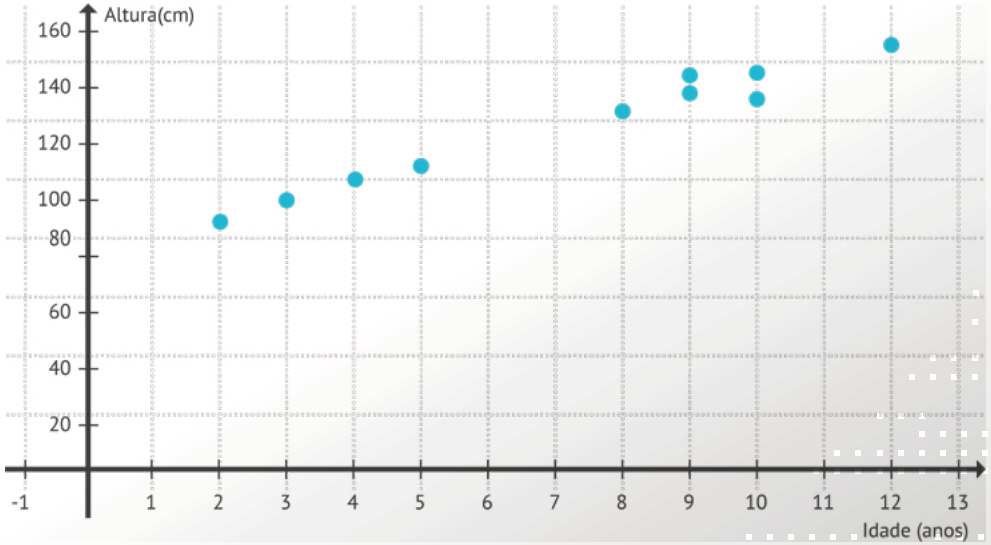


Figura 1: Diagrama de dispersão de idade por altura. Fonte: Elaborada pelo autor.

Podemos perceber, pela disposição dos pontos no plano cartesiano, que a altura dos indivíduos aumenta de acordo com o aumento da idade. Essa relação é esperada, pois sabemos que, com o aumento da idade, as crianças tendem a crescer em altura também. Esse tipo de relação pode

não ser tão intuitiva, e os diagramas de dispersão fornecem uma noção sobre se as variáveis que estamos analisando demonstram algum tipo de relação, assim como o sentido e a intensidade.

Observe os dados fictícios encontrados em Mattos, Azambuja e Konrath (2017), apresentados na próxima tabela, que mostram a idade de 10 indivíduos e o tempo médio de permanência na frente do computador.

Indivíduo	Idade, em anos(x)	Tempo de permanência, em minutos (y)
1	32	290
2	44	150
3	26	340
4	44	100
5	40	130
6	36	180
7	28	290
8	40	200
9	34	220
10	20	380

Figura 2: Idade de 10 indivíduos e o tempo médio de permanência na frente do computador. Fonte: Adaptado de MATTOS, AZAMBUJA e KONRATH, 2017.

O diagrama de dispersão dos dados está representado a seguir.

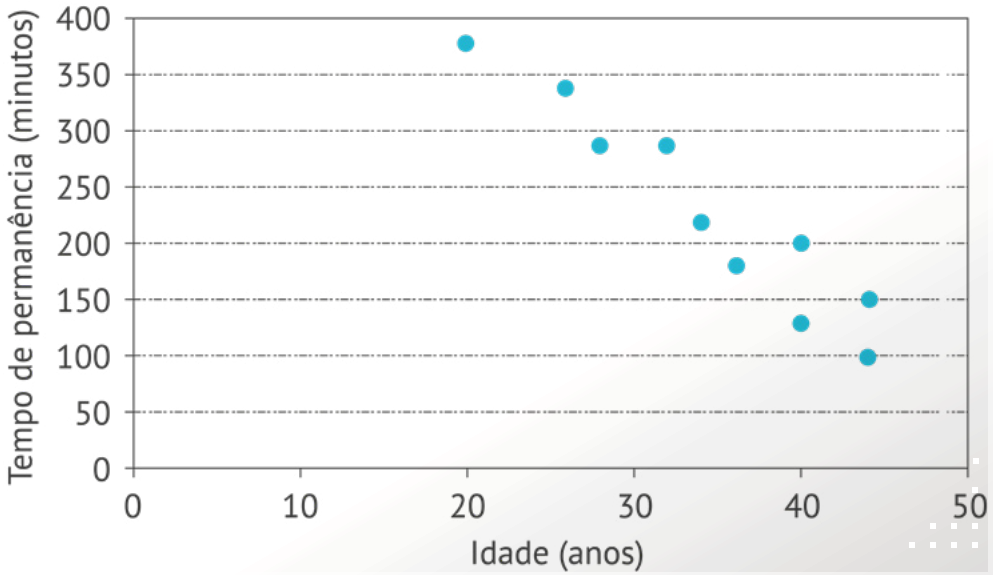
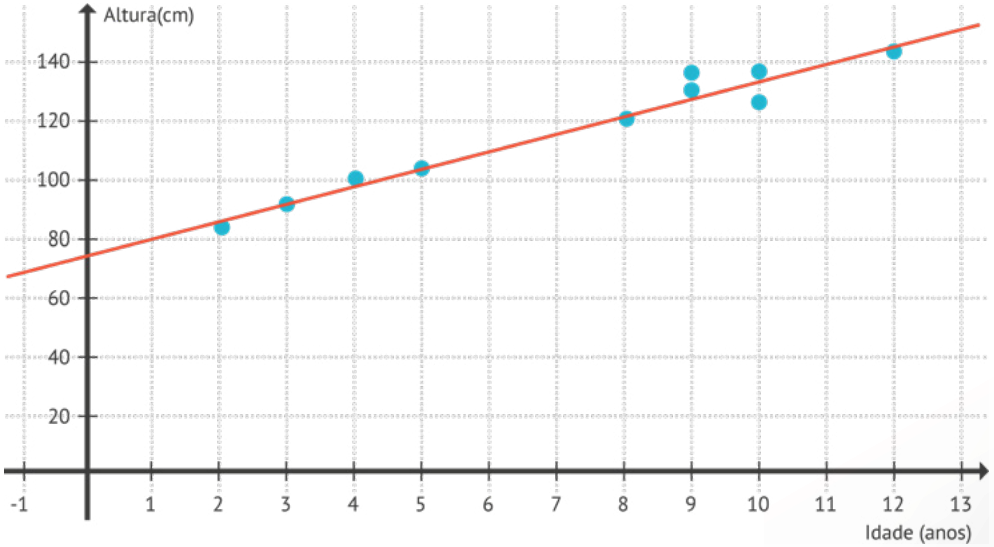


Figura 3: Diagrama de dispersão. Fonte: Adaptado de MATTOS, AZAMBUJA e KONRATH, 2017.

Note que, quanto maior a idade, o tempo médio gasto em frente ao computador diminui. Isso também consegue demonstrar uma relação entre as duas variáveis, mesmo que o sentido da relação seja invertido, ou seja, quando uma variável aumenta, a outra diminui. A relação entre a idade e o tempo de permanência na frente do computador pode não ser tão intuitiva; portanto, o diagrama de dispersão, nesse caso, auxilia a visualização.

Apesar de ser uma ferramenta muito útil para uma primeira análise mais intuitiva da relação entre duas variáveis observáveis em uma amostra, o diagrama de dispersão, por si só, não é suficiente para que possamos estabelecer a correlação entre as variáveis. As relações entre as duas variáveis podem ser diversas, mas, aqui, iremos focar no caso em que essa relação é linear, ou seja, estaremos interessados em estimar a relação entre elas por meio de uma reta. Observe o diagrama de dispersão das observáveis apresentadas na primeira tabela e a reta que melhor aproxima a relação entre elas.



Importante

Existem diversos softwares estatísticos e matemáticos que podem auxiliar na plotagem dos diagramas de dispersão, facilitando a análise de correlação entre duas variáveis. Procure um que atenda melhor às suas necessidades e faça alguns experimentos com ele, para que você se familiarize com a interface e possa extrair o máximo de informações e recursos deste software.

Vamos agora compreender como analisar essa correlação e como determinar a equação da reta que descreve essa relação aproximada.

2. Coeficiente de Correlação Linear de Pearson

Considere duas variáveis x e y que são observadas em pares sob um mesmo indivíduo da nossa população estatística. Estamos interessados em estudar se há correlação entre as variáveis. Nos exemplos anteriores, nossas variáveis foram "idade e altura" e "idade e tempo de permanência em frente ao computador". Considere ainda que estamos avaliando uma coleção de n pares de valores para realizar nosso estudo de correlação, ou seja,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

Considere que $\overline{x}$ e $\overline{y}$ denotam a média aritmética de cada uma das variáveis calculada separadamente. Nessas condições, definimos o **coeficiente de correlação linear de Pearson** ou **coeficiente de correlação de Pearson** como sendo

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Curiosidade

Karl Pearson nasceu em 1857, no Reino Unido, e foi um dos grandes estatísticos com diversas contribuições à área. Ele foi responsável pela criação do Departamento de Estatística Aplicada na University College of London, em 1911, sendo o primeiro departamento com dedicação exclusiva para a Estatística. Pearson tem diversos artigos publicados e dedicou grande parte da vida aos estudos e à docência. Faleceu no dia 27 de abril de 1936, aos 79 anos, e continua sendo uma das grandes referências para a Estatística até hoje.

Karl Person (1857-1936).

Vamos realizar o cálculo desse coeficiente em um exemplo numérico, cujos dados são apresentados na tabela a seguir.

x	y
2	12
4	19
5	24
6	31
8	44

Para realizarmos um estudo da correlação entre as variáveis acima pelo coeficiente de correlação de Pearson, podemos complementar a tabela com algumas colunas que irão facilitar o cálculo deste coeficiente. Como pudemos perceber, o cálculo envolverá as médias aritméticas das duas variáveis separadamente. Denotando por $\overline{x}$ a média da variável x e $\overline{y}$ a média da variável y, obtemos que

$$\overline{x} = \frac{2+4+5+6+8}{5} = \frac{25}{5} = 5$$

e

$$\overline{y} = \frac{12+19+24+31+44}{5} = \frac{130}{5} = 26.$$

As próximas colunas que iremos adicionar à nossa tabela são as diferenças de cada uma das observáveis com a sua respectiva média. Dessa forma, ficamos com uma tabela como a apresentada a seguir.

x	y	$x-\overline{x}$	$y-\overline{y}$
2	12	-3	-14
4	19	-1	-7
5	24	0	-2
6	31	1	5
8	44	3	18

A seguir, calcularemos o quadrado das diferenças separadamente, pois devemos calcular $\sum_{i=1}^{n}(x_i-\overline{x})^2$ e $\sum_{i=1}^{n}(y_i-\overline{y})^2$. Dessa forma, temos uma tabela com seis colunas como mostrado abaixo.

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$
2	12	-3	-14	9	196
4	19	-1	-7	1	49
5	24	0	-2	0	4
6	31	1	5	1	25
8	44	3	18	9	324

Por fim, a última coluna que iremos adicionar à nossa tabela original é a coluna que indicará o produto $(x-\overline{x})\cdot(y-\overline{y})$, obtendo a seguinte tabela.

x	y	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})\cdot(y-\overline{y})$
2	12	-3	-14	9	196	42
4	19	-1	-7	1	49	7

5	24	0	-2	0	4	0
6	31	1	5	1	25	5
8	44	3	18	9	324	54

Baseado nos valores encontrados, calcularemos os somatórios que fazem parte do cálculo do coeficiente de correlação de Pearson:

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = 9 + 1 + 0 + 1 + 9 = 20,$$

$$\sqrt{20} \approx 4,4721,$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = 196 + 49 + 4 + 25 + 324 = 598,$$

$$\sqrt{598} \approx 24,454,$$

$$\sqrt{20} \cdot \sqrt{598} \approx 109,36$$

e

$$\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = 42 + 7 + 0 + 5 + 54 = 108.$$

Portanto, o coeficiente de correlação linear de Pearson para essas variáveis é de

$$r = \frac{108}{109,36} = 0,9875.$$

É importante destacar que o coeficiente de correlação linear de Pearson sempre irá variar entre -1 e 1. Quanto maior for o módulo deste coeficiente, maior será a correlação entre as duas variáveis analisadas. Utilizando os parâmetros definidos em Mattos, Azambuja e Konrath (2017), classificaremos o grau de correlação a partir do coeficiente linear de Pearson segundo as seguintes regras:

- se $|r| \leq 0,5$, então, a correlação entre as variáveis é fraca;
- se $0,5 < |r| < 0,8$, então, a correlação entre as variáveis é média ou moderada; e
- se $|r| \geq 0,8$, então, a correlação entre as variáveis é forte.

O sinal do coeficiente de correlação linear de Pearson indica se as variáveis são diretamente proporcionais, caso o sinal seja positivo, ou inversamente proporcionais, caso o sinal seja negativo.

Importante

Não se esqueça de que o grau de correlação entre as variáveis analisadas está relacionado com o módulo do coeficiente de Pearson. O sinal apenas indica se as grandezas analisadas são diretamente ou inversamente proporcionais.

No exemplo anterior, podemos afirmar que há uma correlação linear forte entre as variáveis e que ela é diretamente proporcional. Veja o diagrama de dispersão das variáveis x e y.

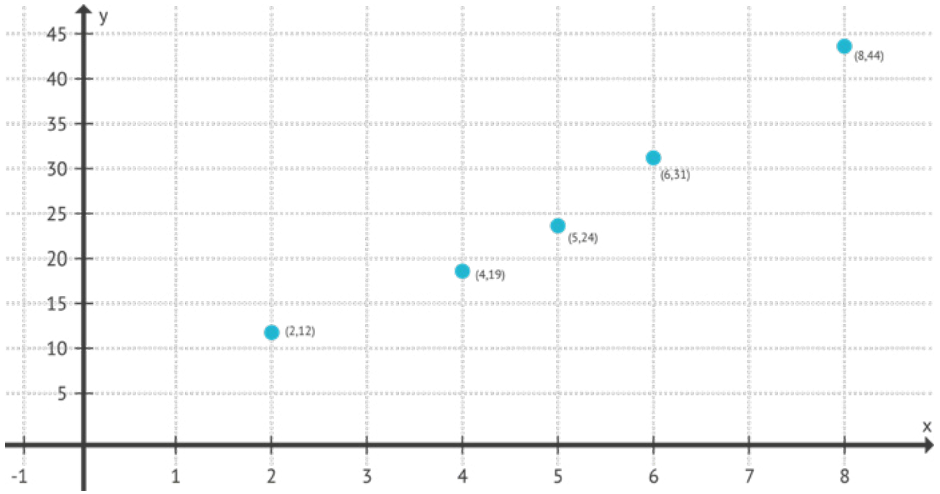


Figura 4: Gráfico de dispersão das variáveis x e y. Fonte: Elaborada pelo autor.

3. Equação de Regressão

Para finalizar a análise da correlação linear entre duas variáveis, iremos agora descrever como calcular a equação da reta que melhor aproxima a relação entre as variáveis estudadas. Considere duas variáveis x e y que são observadas em n pares

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

e $\overline{x}$ e $\overline{y}$ as médias aritméticas de cada uma das variáveis calculadas separadamente. O procedimento que iremos descrever aqui é conhecido como **regressão linear**. Vamos representar a relação entre as variáveis por uma equação do tipo

$$y = ax + b$$

onde a é o coeficiente angular da reta e b o coeficiente linear. Determinando esses dois valores, temos bem definida a equação que modela a relação entre x e y. Os valores são calculados segundo as fórmulas

$$a = \frac{\sum_{i=1}^{n} (x - \overline{x}) \cdot (y - \overline{y})}{\sum_{i=1}^{n} (x - \overline{x})^2}$$

e $b = \overline{y} - a \cdot \overline{x}$.

No exemplo apresentado para exemplificar o cálculo do coeficiente de correlação linear de Pearson na seção anterior, podemos calcular facilmente os coeficientes da reta que melhor estima a relação entre as variáveis a partir da última tabela construída. Temos que

$$\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = 42 + 7 + 0 + 5 + 54 = 108$$

e

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = 9 + 1 + 0 + 1 + 9 = 20,$$

portanto

$$a = \frac{\sum_{i=1}^{n}(x - \overline{x}) \cdot (y - \overline{y})}{\sum_{i=1}^{n}(x - \overline{x})^2} = \frac{108}{20} = 5,468$$

e

$$b = \overline{y} - a \cdot \overline{x} = 26 - 5,468 \cdot 5 = -1,34.$$

A reta que melhor expressa a relação entre as variáveis x e y neste caso é

$$y = 5,468 \cdot x - 1,34.$$

Observe o diagrama de dispersão dos pontos considerados e o gráfico da reta da equação encontrada.

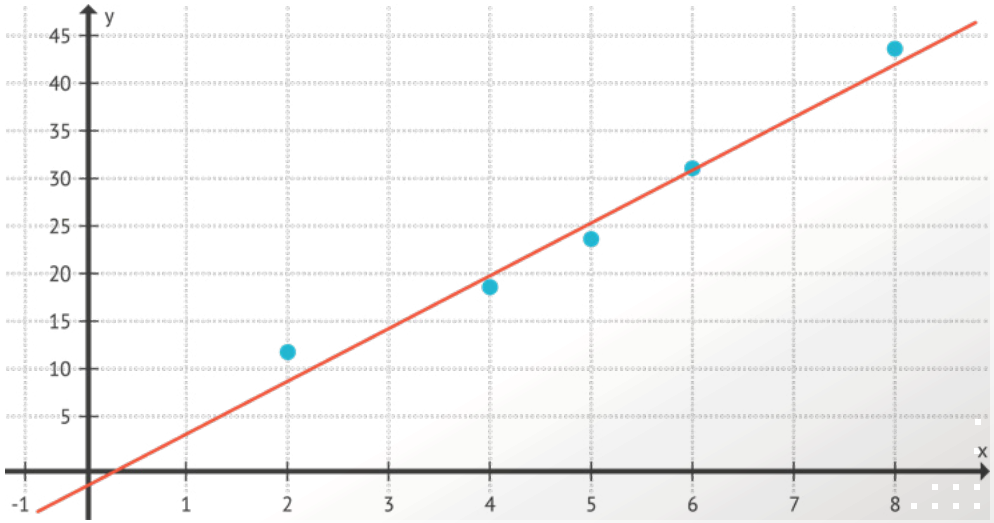


Figura 5: Diagrama de dispersão e gráfico da reta encontrada pela regressão linear.
Fonte: Elaborada pelo autor.

Vejamos outro exemplo do cálculo do coeficiente de correlação linear de Pearson e do procedimento de regressão linear.

Exemplo 1: considere os dados apresentados na tabela que relaciona o tempo médio de permanência em frente ao computador de acordo com a idade dos indivíduos avaliados, apresentada na primeira seção desta unidade e, novamente, a seguir.

Idade	Tempo (min)
32	290
44	150
26	340
44	100
40	130
36	180
28	290
40	200
34	220
20	380

Identificando a variável idade por x e a variável do tempo médio de permanência em frente ao computador por y, podemos calcular suas médias $\overline{x}$ e $\overline{y}$ como apresentado a seguir.

$$\overline{x} = \frac{32+44+26+44+40+36+28+40+34+20}{10} = \frac{344}{10} = 34,4$$

$$\overline{y} = \frac{290+150+340+100+130+180+290+200+220+380}{10} = \frac{2280}{10} = 228$$

Inserindo novas colunas na tabela apresentada, da mesma forma que procedemos anteriormente, obtemos os seguintes valores:

Idade	Tempo (min)	$x - \overline{x}$	$y - \overline{y}$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$(x-\overline{x})\cdot(y-\overline{y})$
32	290	-2,4	62	5,76	3844	-148,8
44	150	9,6	-78	92,16	6084	-748,8
26	340	-8,4	112	70,56	12544	-940,8
44	100	9,6	-128	92,16	16384	-1228,8
40	130	5,6	-98	31,36	9604	-548,8
36	180	1,6	-48	2,56	2304	-76,8
28	290	-6,4	62	40,96	3844	-396,8
40	200	5,6	-28	31,36	784	-156,8
34	220	-0,4	-8	0,16	64	3,2
20	380	-14,4	152	207,36	23104	-2188,8

Podemos calcular o coeficiente de correlação linear de Pearson para essas variáveis de acordo com as observáveis apresentadas. Observe os cálculos abaixo:

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = 5,76+92,16+70,56+92,16+31,36+2,56+40,96+31,36+0,16+207,36 = 574,4,$$

$$\sqrt{574,4} \approx 23,966\,,$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = 3844 + 6084 + 12544 + 16384 + 9604 + 2304 + 3844 + 784 + 64 + 23104 = 78560\,,$$

$$\sqrt{78560} \approx 280,285\,,$$

$$\sqrt{574,4} \cdot \sqrt{78560} \approx 6717,31,$$

$$\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y}) = -148,8 - 748,8 - 940,8 - 1228,8 - 548,8 - 76,8 - 396,8 - 156,8 + 3,2 - 2188,8 = -6432$$

e, portanto,

$$r = \frac{-6432}{6717,31} = -0,9575\,.$$

A correlação linear entre as duas variáveis estudadas é forte, mas é inversamente proporcional, ou seja, quanto maior a idade dos participantes, menos tempo eles passam em frente à tela do computador diariamente.

Determinado os coeficientes da reta que melhor descreve a relação entre as variáveis analisadas, obtemos:

$$a = \frac{\sum_{i=1}^{n}(x - \overline{x}) \cdot (y - \overline{y})}{\sum_{i=1}^{n}(x - \overline{x})^2} = \frac{-6432}{574,4} \approx -11,2$$

e

$$b = \overline{y} - a \cdot \overline{x} = 228 - (-11,2) \cdot 33,4 = 613,28\,.$$

Portanto, a equação da reta é

$$y = -11,2 \cdot x + 613,28\,.$$

Observe o diagrama de dispersão dos pontos considerados e o gráfico da reta que melhor descreve a relação entre as variáveis.

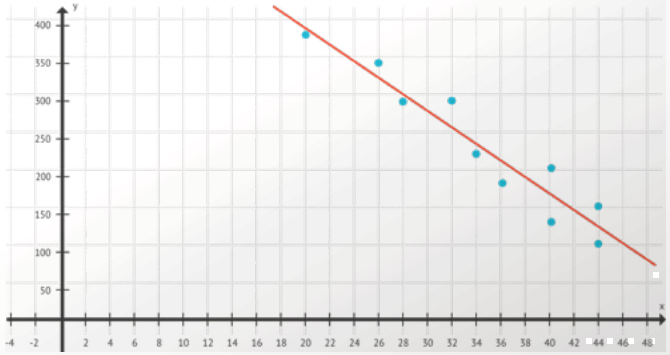


Figura 6: Diagrama de dispersão e gráfico da reta obtido pela regressão linear.
Fonte: Elaborada pelo autor.

O coeficiente de correlação linear de Pearson baseia-se em pares de observáveis e indica se há uma relação linear entre as variáveis consideradas. Os diagramas de dispersão nos auxiliam como um recurso

gráfico, que permite uma análise inicial da possível relação entre nossas variáveis. Quanto maior o número da nossa amostra, maior será a chance de podermos identificar uma relação entre as variáveis. Aprendemos a calcular o coeficiente de Pearson e também como calcular os coeficientes angular e linear da reta que melhor relaciona duas variáveis a partir de um conjunto finito de observações. Esse tipo de análise é muito comum em Estatística, e trata-se de um dos primeiros casos de regressão e correlação que são estudados nessa ciência.

Referências

FONSECA, J. S.; MARTINS, G.A. **Curso de estatística**. 6. ed. São Paulo: Atlas, 2012.

MATTOS, V. L. D.; AZAMBUJA, A. M. V.; KONRATH, A. C. **Introdução à estatística**: aplicações em ciências exatas. Rio de Janeiro: LTC, 2017.

MOORE, D.; NOTZ, W.; FLIGNER, M. **A estatística básica e sua prática**. 7. ed. Rio de Janeiro: LTC, 2017.

MORETTIN, P.; BUSSAB, W. **Estatística básica**. 9. ed. São Paulo: Saraiva, 2017.

PEARSON, K. Statistician: Blue Plaques. English Heritage. **English Heritage**. Disponível em: https://www.english-heritage.org.uk/visit/blue-plaques/karl-pearson/. Acesso em: 14 ago. 2021.

SPIEGEL, M.; SCHILLIER, J.; SRINIVASAN, A. **Probabilidade e estatística**. 3. ed. Porto Alegre: Bookman, 2013.

SPIEGEL, M.; STEPHENS, L. **Estatística**. Porto Alegre: Bookman, 2009.

TRIOLA, M. **Introdução à estatística**. 12. ed. Rio de Janeiro: LTC, 2017.

VIEIRA, S. **Fundamentos de estatística**. 6. ed. São Paulo: Atlas, 2019.