



Cruzeiro do Sul
Virtual
Educação a Distância

REGRESSÃO

Prof. Ismar Frango



Regressão

Uma **reta de regressão** é uma linha reta que descreve como uma **variável resposta y** muda quando uma **variável explicativa x** muda, quando acreditamos que a relação entre y e x é **linear**. Assim, é possível prever, de certa forma o valor de y de acordo com o valor de x

Uma reta que relacione essas duas variáveis tem uma equação da forma:

$$y = a + bx$$

Na realidade, esta equação serve para descrever **qualquer** reta.

Nesta equação:

a é o **intercepto** (**coeficiente linear**)

Indica o valor de y quando $x=0$.

b é a **inclinação** (**coeficiente angular**)

Indica o quanto y muda de acordo com as mudanças em x.



Em alguns livros de Matemática, essa equação aparece com os coeficientes a e b com papéis trocados. É comum encontrar essa fórmula como:

$$y = ax + b$$

Por isso, sempre que possível, usamos o termo inclinação (ou **coeficiente angular**) para indicar o coeficiente que multiplica x, e intercepto (ou **coeficiente linear**) para o coeficiente independente.

Vejamos o exemplo a seguir:

Desde dezembro de 2018, a tarifa de táxi comum na cidade do Rio de Janeiro é assim calculada: o valor da **bandeirada** (valor inicial da corrida) passou a **R\$ 5,80** e o **valor por km** rodado é de **R\$ 2,60** na bandeira 1 (*praticada das 6h às 21h, de segunda a sábado*), e de **R\$ 3,12** na bandeira 2 (*que vale para as viagens noturnas, das 21h às 6h, de segunda a sábado, domingos, feriados e subidas íngremes a qualquer horário*)

Podemos dizer que as corridas de táxi no Rio seguem a seguinte função:

$$\text{valor} = \text{bandeirada} + \text{valor_km} * \text{km_rodado}$$

Para a bandeira 1, a função é

$$y = 5,80 + 2,60x$$

Já para a bandeira 2, a função é

$$y = 5,80 + 3,12x$$

Onde y é o valor final em R\$ e x é a quantidade de km rodados.

Assim, uma corrida de 10km na bandeira 1 custaria:

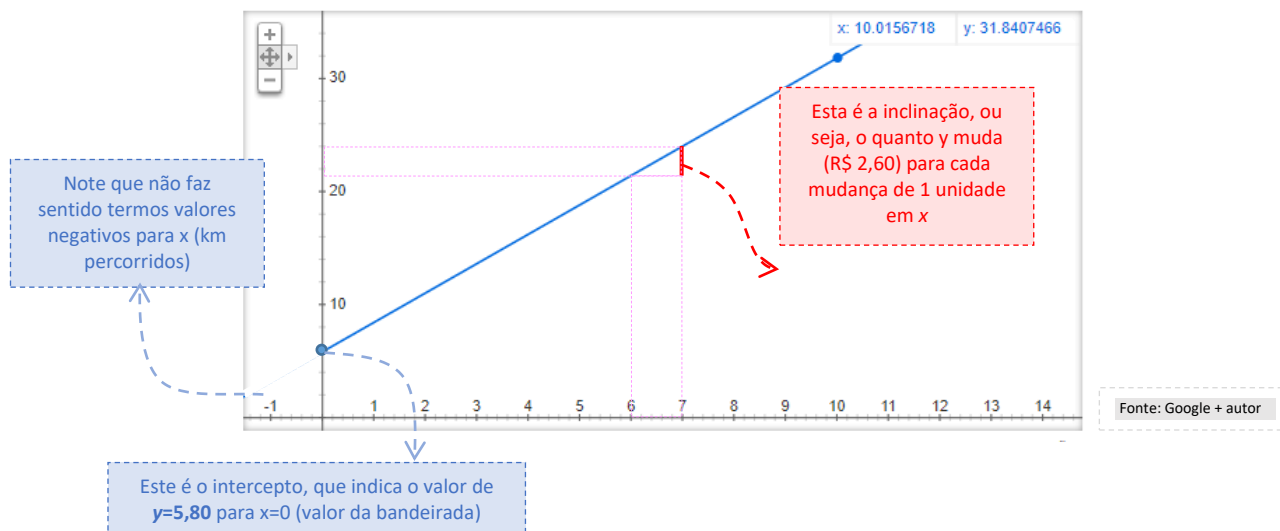
$$y = 5,80 + 2,60 * 10 = 31,80$$

Se fosse na bandeira 2, a mesma corrida custaria:

$$y = 5,80 + 3,12 * 10 = 37,00$$

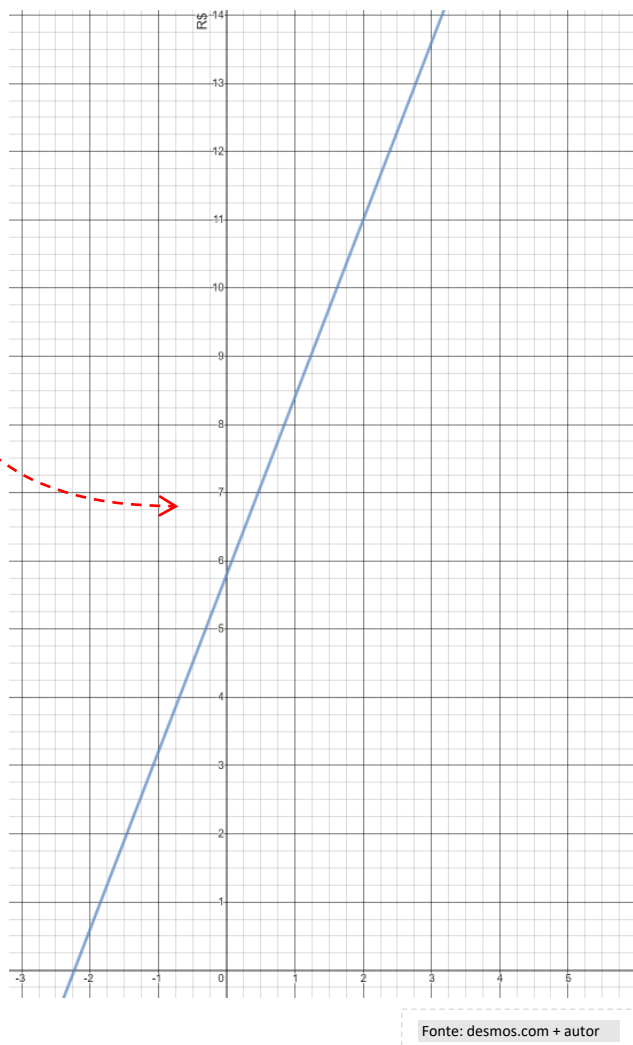
O gráfico da reta de parte do exemplo anterior (Bandeira 1) é mostrado a seguir (o gráfico para Bandeira 2 seria parecido, exceto pela inclinação da reta).

Gráfico para $5.8+2.6*x$



Atenção: o gráfico acima, por razões didáticas, adota escalas diferentes para os eixos x e y. Caso fôssemos adotar a mesma escala para os dois eixos, a aparência do gráfico seria **assim**.

Note que qualquer alteração na escala afeta visualmente a inclinação da reta.

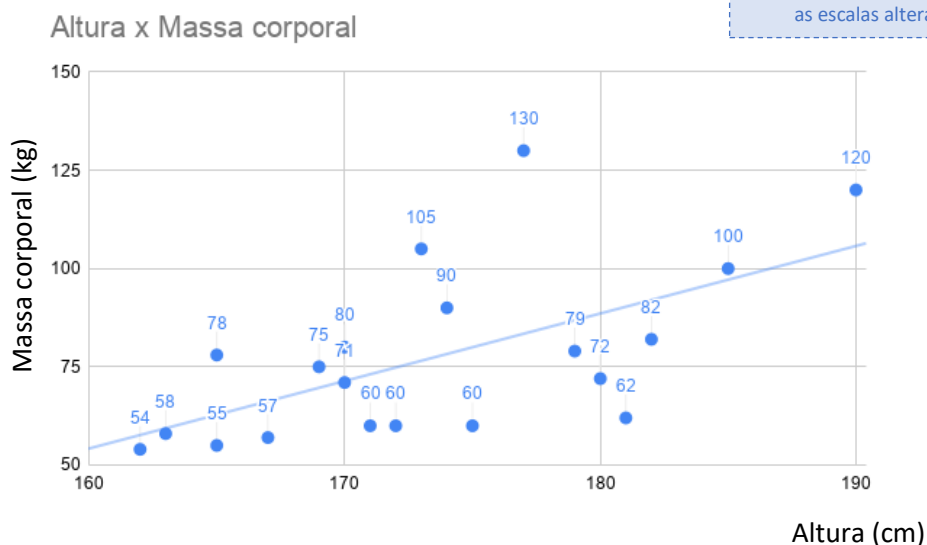


O exemplo dado mostra o caso de uma relação entre x (km rodados) e y (valor a pagar) em que o valor de y é resultado direto da aplicação. Porém, em muitas situações reais, y não é resultado direto de alguma função linear aplicada a x . Dessa forma, os pontos de um gráfico representando uma distribuição raramente ficarão exatamente na reta – ao contrário, estão, quando muito, próximos de uma reta calculada a partir desses pontos.

Dessa maneira, você pode estar se perguntando: como calcular uma reta que melhor se **aproxime** de dados **reais** de observações?

Vejamos o seguinte exemplo, já trabalhado em outras unidades, que envolve o *dataset* de massas corporais e alturas. Apresentamos o diagrama de dispersão da relação altura (x) por massa corporal (y). Veja que incluímos uma reta que se aproxima dos dados reais, bem como os valores exatos de massa corporal para cada ponto. Note que, à exceção do ponto (170;71), nenhum outro ponto está **exatamente** sobre a reta.

Altura	Massa corporal
180	72
170	80
175	60
174	90
185	100
190	120
182	82
179	79
165	78
165	55
170	71
169	75
177	130
173	105
172	60
162	54
163	58
167	57
171	60
181	62

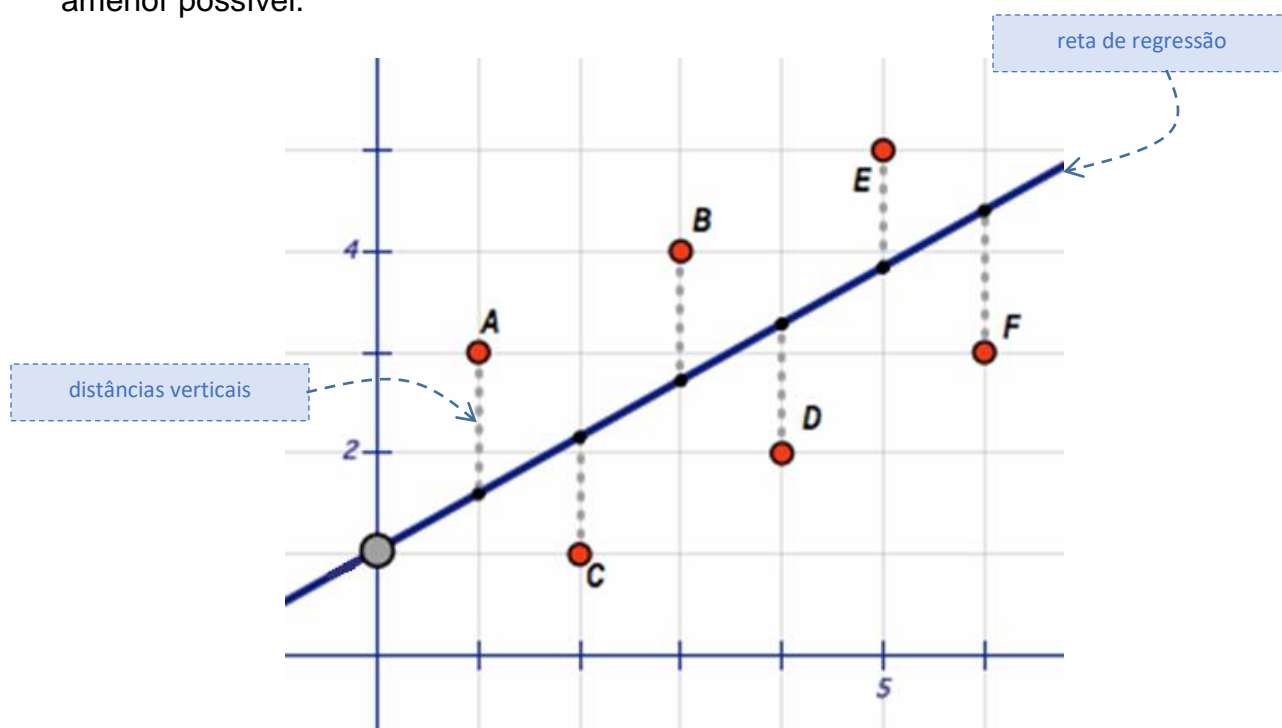


Método dos mínimos quadrados

Temos que encontrar algum meio de chegar à equação de uma reta que seja uma boa estimativa dos dados presentes em uma distribuição. Uma das maneiras de fazer isso é tentando minimizar o **erro da previsão** feita pela reta de regressão, fazendo com que a **distância, na vertical, entre os pontos observados e a reta, seja a menor possível**. Este é o Método dos Mínimos Quadrados.

A **reta de regressão de mínimos quadrados** é a reta que minimiza soma dos **quadrados das distâncias** verticais entre os pontos observados à reta.

Veja o seguinte exemplo: sejam cinco pontos A, B, C, D, E e F, a reta de regressão calculada pelo método dos mínimos quadrados é a reta que faz com que as distâncias verticais de cada um dos pontos em relação a essa reta seja amenor possível.



Fonte: Wikimedia Commons (Licença CC)

Como construir essa reta?

Considere um *dataset* com os dados de uma **variável explicativa x** e de uma **variável resposta y** para n indivíduos.

A reta de regressão de mínimos quadrados é a reta:

lê-se "y chapéu"

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

Onde:

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

$$b = r \frac{s_y}{s_x}$$

Ou seja:

O **intercepto a** (coeficiente linear) é dado pela média de y (\bar{y}) menos a inclinação b (coeficiente angular) vezes a média de x (\bar{x}).

A **inclinação b** (coeficiente angular) é dada pela correlação (r) vezes o desvio-padrão de x (s_x) dividido pelo desvio-padrão de y (s_y).

Vejamos no nosso exemplo anterior:

Variável	Média	Desvio-padrão	Correlação	0,5979
x – Altura (cm)	$\bar{x}=173,5$	$s_x = 7,6192$		
y – Massa corporal (kg)	$\bar{y} = 77,4$	$s_y = 21,8930$		

Use o desvio-padrão amostral – nas planilhas de cálculo, use a função DESVPAD.A ou STDEV.S

Logo,

$$b = r \frac{s_y}{s_x} = 0,5979 * \frac{21,893}{7,6192} = 1,718$$

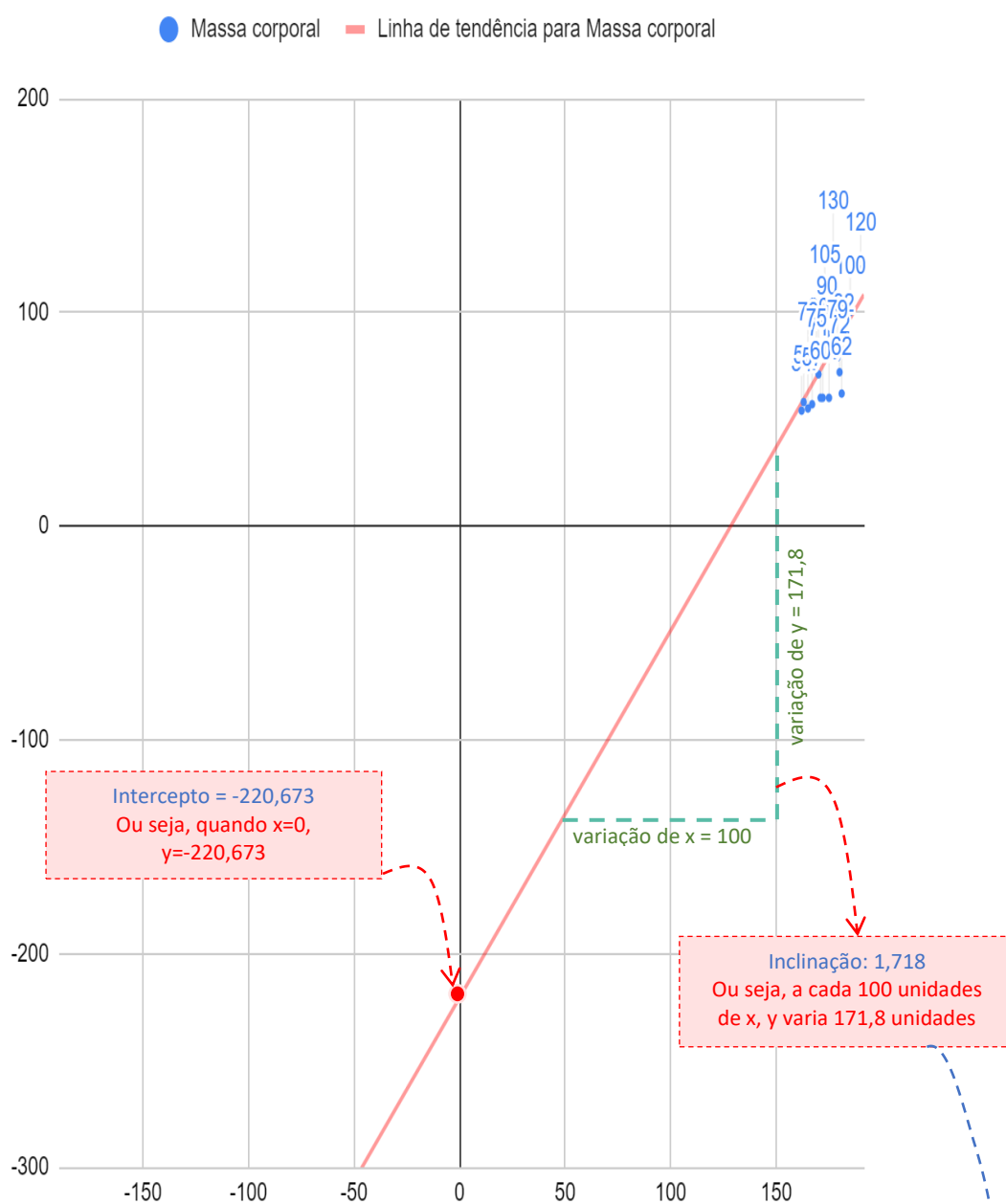
$$a = \bar{y} - b\bar{x} = 77,4 - 1,718 * 173,5 = -220,673$$

$$\hat{y} = a + bx \rightarrow \hat{y} = -220,673 + 1,718x$$

A reta dos mínimos quadrados e o gráfico de dispersão

Vamos rerepresentar o gráfico de dispersão da relação altura x massa corporal, mas agora mantendo uma escala 1:1:

Altura x Massa corporal



Fonte: autor

Sendo ainda mais precisos, uma mudança de 1 desvio-padrão em x corresponde a uma mudança de r desvios-padrões em y

Como pôde ser visto no gráfico anterior, a visualização dos pontos da distribuição ficou bastante comprometida na escala 1:1 adotada. Na verdade, mostramos esse gráfico apenas para você situar melhor a localização do intercepto e compreender o significado da inclinação.

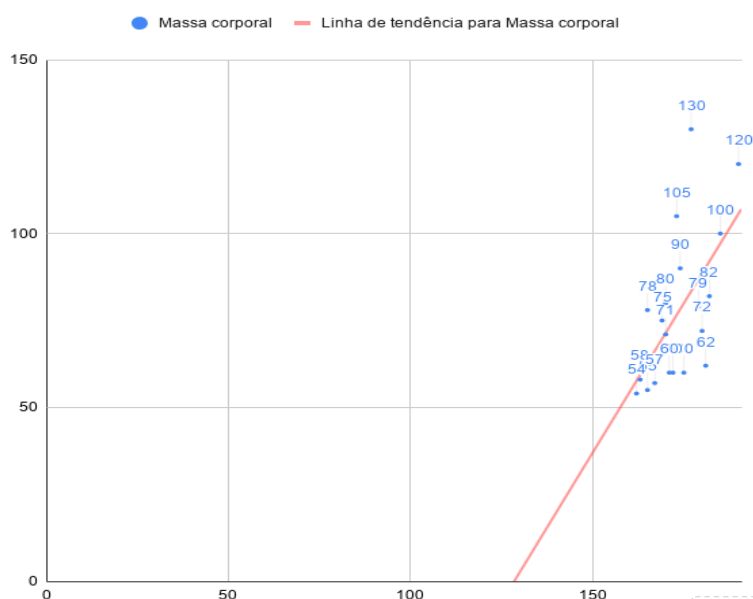


Fonte: Maxpixel (Licença CC0)

Seguindo a reta de regressão dos mínimos quadrados, muitos valores não fazem o menor sentido no mundo real: por exemplo, a localização do intercepto na ponto (0;-220,673) indica que, caso existisse uma pessoa com altura zero, ela deveria pesar aproximadamente 220 “quilos negativos”, o que obviamente não existe.

Ainda que nos concentrássemos apenas no primeiro quadrante (apenas valores positivos; afinal, não existem alturas nem massas corporais negativas em nosso contexto), mantendo a escala 1:1, a visualização do diagrama de dispersão ainda não seria a mais adequada. Observe:


Altura x Massa corporal



Fonte: autor

Isso se dá por uma razão muito simples: dados reais em geral são concentrados em um intervalo específico. Por isso é importante conhecermos a amplitude e o resumo de cinco números de uma distribuição (a saber, o valor mínimo, os três quartis - o que inclui a mediana – e o valor máximo) Por exemplo, no caso de nosso *dataset*, as alturas variam aproximadamente entre 160 e 190 cm, enquanto as massas corporais não saem do intervalo de 50 a 130kg. Por isso, é mais adequada a representação gráfica desse recorte do plano cartesiano onde se concentram os dados ($160 \leq x \leq 190$; $50 \leq y < 130$), sempre identificando claramente que se adotou uma visualização com escala alterada.

Vamos trabalhar com um outro exemplo:



°C	Vendas
14,2	R\$ 215
16,4	R\$ 325
11,9	R\$ 185
15,2	R\$ 332
18,5	R\$ 406
22,1	R\$ 522
19,4	R\$ 412
25,1	R\$ 614
23,4	R\$ 544
18,1	R\$ 421
22,6	R\$ 445
17,2	R\$ 408

As temperaturas no outono de 2019 em Juiz de Fora - MG oscilaram bastante. Juiz de Fora é uma cidade de clima ameno e temperatura instável, com picos de frio ou calor na mesma semana, em especial no outono – é uma época boa para coletar dados, dada a variação de temperatura.

O *dataset* ao lado mostra a variação de temperatura durante 12 dias do outono de 2019, bem como as vendas de sorvete em uma popular sorveteria da cidade.

A partir desses dados, é possível fazer uma previsão de vendas de sorvete considerando a temperatura de um dia

Vamos resolver isso calculando a reta de regressão pelo método dos mínimos quadrados, como vimos anteriormente!

Vamos lá! A primeira ação a ser tomada é compreender quem será a **variável explicativa x** e quem será a **variável resposta y**. Embora isso não tenha nenhum impacto para o cálculo da correlação r , na definição da reta de regressão isso é de extrema importância.

Isso, obviamente, depende da contexto em que foram obtidos os dados, e o que eles significam. Em nosso caso, tem sentido imaginarmos que as vendas de sorvete serão impactadas pelo aumento ou diminuição da temperatura – logo, a **variável explicativa x seria a temperatura**, enquanto a **variável resposta y seria a venda de sorvete**. Não faz sentido imaginarmos o contrário – que as vendas de sorvete fazem a temperatura subir ou descer.

Isto posto, vamos calcular os valores das médias e desvios-padrão de ambas as variáveis e também a correlação entre elas.

Variável	Média	Desvio-padrão
x – Temperatura (°C)	$\bar{x}=18,675$	$s_x = 4,0111$
y – Vendas de sorvete (R\$)	$\bar{y} = 402,42$	$s_y = 126,0429$

Correlação	0,9575
-------------------	---------------

Usamos o desvio-padrão amostral – nas planilhas de cálculo, é a função DESVPAD.A ou STDEV.S

Veja que é uma correlação positiva forte

Logo,

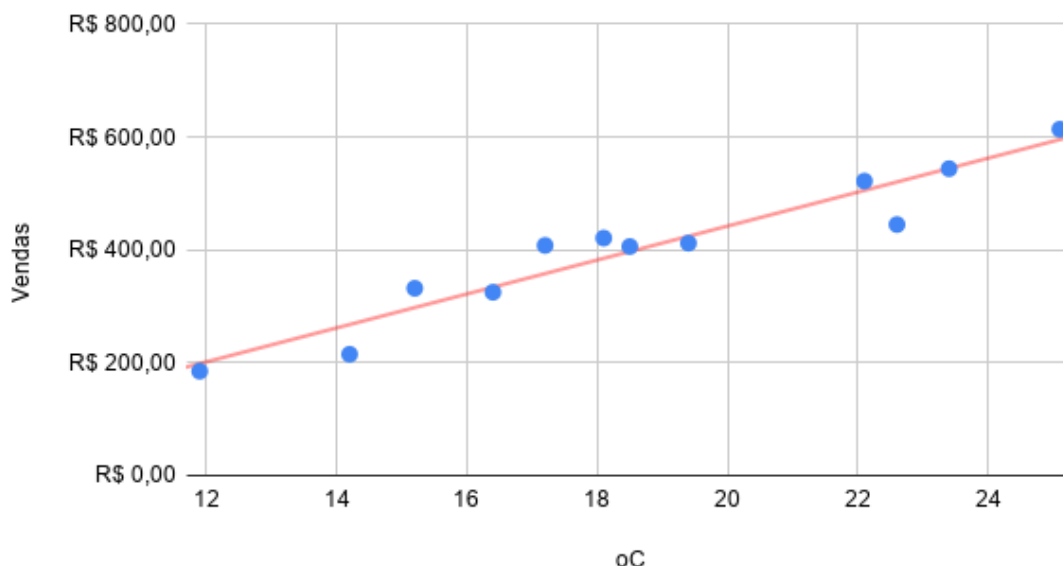
$$b = r \frac{s_y}{s_x} = 0,9575 * \frac{126,0429}{4,0111} = 30,088$$

$$a = \bar{y} - b\bar{x} = 402,42 - 30,088 * 18,675 = -159,4734$$

$$\hat{y} = a + bx \rightarrow \hat{y} = -159,4734 + 30,088x$$

A reta de regressão dos mínimos quadrados pode ser vista no gráfico de dispersão a seguir:

Vendas x oC



Fonte: autor

Com a equação da reta de regressão, é possível fazer previsões sobre a venda de sorvetes, de acordo com a variação da temperatura. Por exemplo:

Qual será a venda de sorvetes se a temperatura atingir os 40°C?

$$\hat{y} = -159,4734 + 30,088x \rightarrow -159,4734 + 30,088 * 40 = R\$1044,05$$

Qual será a venda de sorvetes se a temperatura atingir os 8°C?

$$\hat{y} = -159,4734 + 30,088x \rightarrow -159,4734 + 30,088 * 8 = R\$81,23$$

Abaixo de qual temperatura nem adianta abrir a sorveteria (vendas=0)?

$$\hat{y} = -159,4734 + 30,088x \rightarrow -159,4734 + 30,088 * x = 0 \rightarrow$$

$$x = \frac{159,4734}{30,088} = 5,3^{\circ}C$$

Ou seja, abaixo de 5,3°C, ninguém em Juiz de Fora anima a encarar um sorvetinho.



Para saber mais, leia o capítulo 5 do e-book:

MOORE, David S.; NOTZ, William I.; FLINGER, Michael A. A
Estatística Básica e sua Prática. 7 ed. Rio de Janeiro: LTC,
2017 – Capítulo 5