Universidade Federal do Paraná
Campus Avançado de Jandaia do Sul
Estatística

Correlação e Regressão

Dr. Landir Saviniec

E-mail: landir.saviniec@gmail.com
Homepage: github.com/lansaviniec

Outubro de 2018

Visão geral

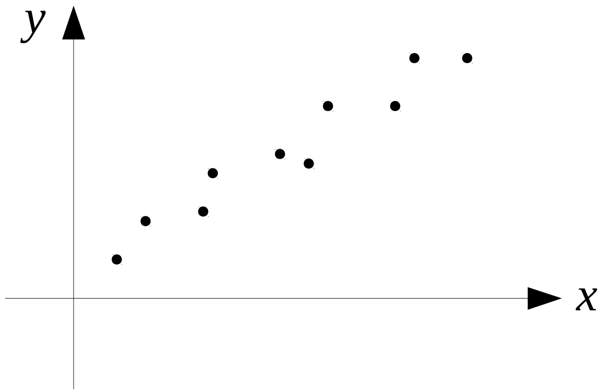
Em **correlação e regressão linear**, analisamos se duas variáveis se **correlacionam de forma linear** e se podemos **predizer o valor de uma, em função da outra**, **por** meio de **uma função linear**.

Correlação

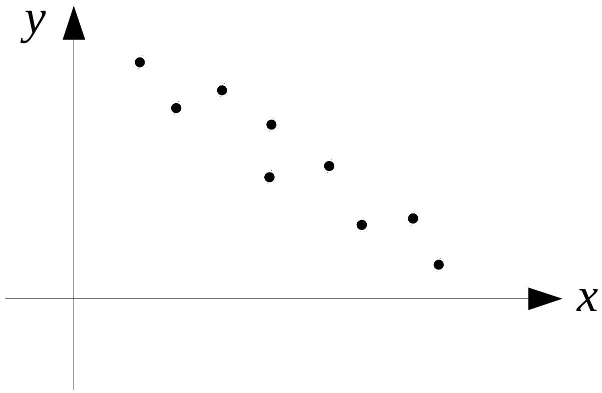
Correlação

Existe uma **correlação entre duas variáveis quando os valores de uma variável estão relacionadas**, de alguma maneira, **com os valores da outra variável.**

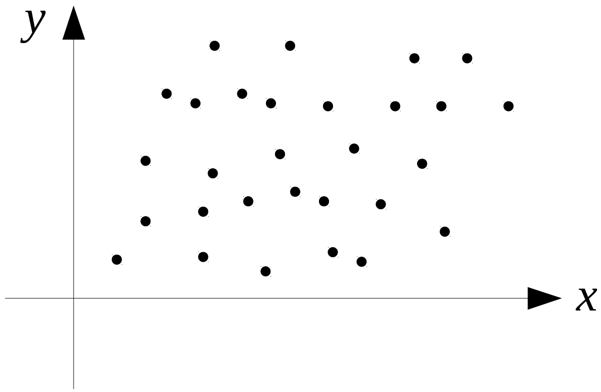
Exemplo



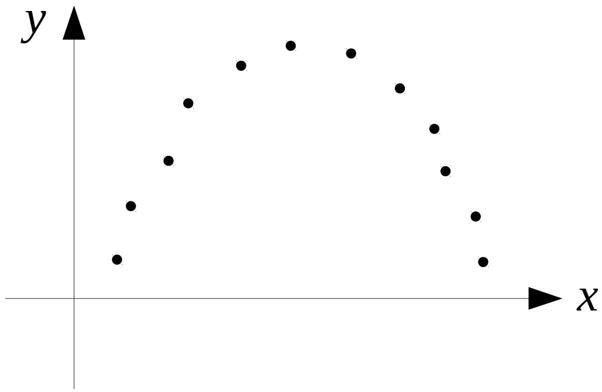
Correlação linear positiva



Correlação linear negativa



Nenhuma correlação



Correlação não-linear

Coeficiente de correlação linear

O **coeficiente de correlação linear r** mede a força da correlação linear entre valores quantitativos emparelhados x e y em uma amostra.

Objetivo

Determinar se existe uma correlação linear entre duas variáveis.

Notação

n : *tamanho da amostra de dados emparelhados* $(x\ ,\ y)$

r : *coeficiente de correlação linear amostral*

ρ : *coeficiente de correlação linear populacional*

Requisitos

1) A amostra é uma amostra aleatória simples.
2) O gráfico de dispersão (x,y) se aproxima de uma reta.
3) Valores atípicos devem ser removidos caso se saiba que são erros.

Fórmula para cálculo de r

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

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Ou por:

$$r = \frac{\sum \left(z_x z_y\right)}{n-1}$$

Onde:

$$z_x = \frac{x - \bar{x}}{s_x} \qquad e \qquad z_y = \frac{y - \bar{y}}{s_y}$$

z_x: *escore z para a variável x, cuja média e o desvio padrão amostral são* $\bar{x}$ *e* s_x

z_y: *escore z para a variável y, cuja média e o desvio padrão amostral são* $\bar{y}$ *e* s_y

Propriedades do coeficiente de correlação linear

1) O valor de r está ente -1 e 1. Isto é: $-1 \leq r \leq 1$

2) O valor de r não muda se todos os valores de qualquer das variáveis forem convertidos para uma escala diferente.

3) O valor de r é sensível a valores atípicos.

Exemplo 1

A tabela a seguir apresenta custos de uma fatia de pizza e tarifas de metrô (em dolares) em diferentes meses. Calcule o coeficiente de correlação linear entre as duas variáveis.

Custo da pizza (x)	Custo da tarifa de metrô (y)
0.15	0.15
0.35	0.35
1.00	1.00
1.25	1.35
1.75	1.50
2.00	2.00

Exemplo 1

A tabela a seguir apresenta custos de uma fatia de pizza e tarifas de metrô em diferentes meses. Calcule o coeficiente de correlação linear entre as duas variáveis.

Pizza (x)	Metrô (y)	x^2	y^2	xy
0.15	0.15	0.0225	0.0225	0.0225
0.35	0.35	0.1225	0.1225	0.1225
1.00	1.00	1.0000	1.0000	1.0000
1.25	1.35	1.5625	1.8225	1.6875
1.75	1.50	3.0625	2.2500	2.6250
2.00	2.00	4.0000	4.0000	4.0000
$\sum$: 6.50	6.35	9.7700	9.2175	9.4575

$$r = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}} = \frac{6(9.4575) - (6.50)(6.35)}{\sqrt{6(9.77) - (6.50)^2}\sqrt{6(9.2175) - (6.35)^2}} = \frac{15.47}{\sqrt{16.37}\sqrt{14.9825}} = 0.988$$

Teste de hipótese para o coeficiente de correlação linear

Utilizamos a **estatística de teste t** baseada na **distribuição de Student,** com **n-2 graus de liberdade**:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad com\ GL = n-2$$

Exemplo 2: testando se há correlação linear significante entre as variáveis

Teste se há uma correlação linear significante entre os custos de uma fatia de pizza e as tarifas de metrô do Exemplo 1. Use um nível de significância de 5%.

Passo 1: Formulamos as hipóteses nula e alternativa:

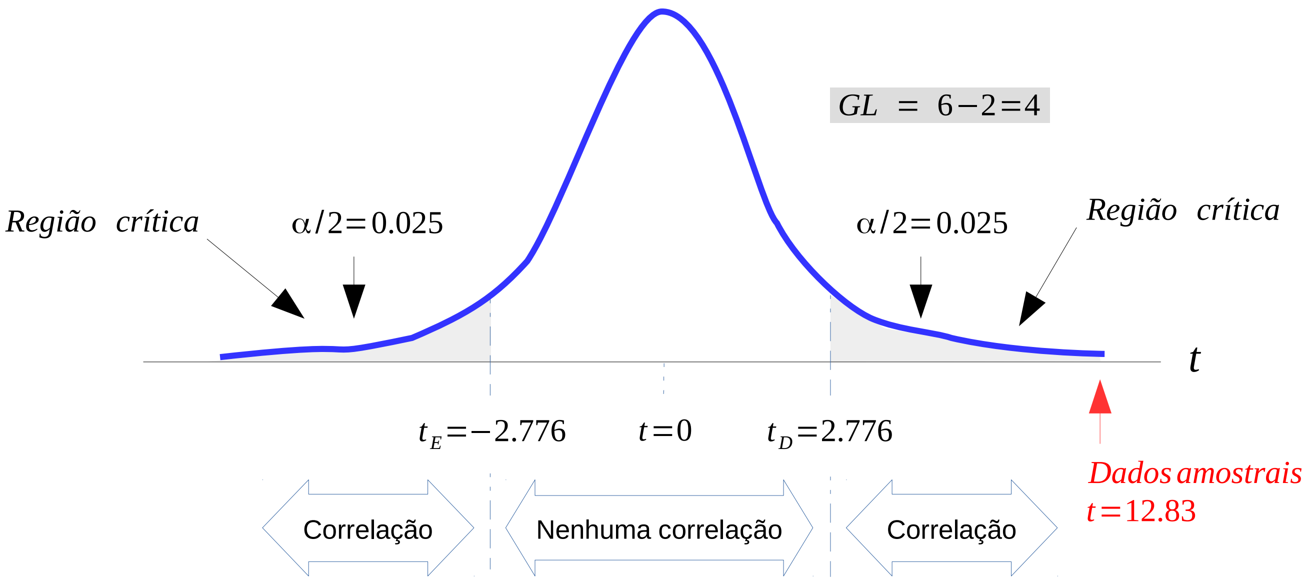
$$H_0: \quad \rho = 0 \quad \left(\textit{não há nenhuma correlação}\right)$$
$$H_1: \quad \rho \neq 0 \qquad \left(\textit{há uma correlação linear}\right)$$

Passo 2: Calculamos a estatística de teste t para os dados amostrais:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.988\sqrt{6-2}}{\sqrt{1-0.988^2}} = \frac{1.976}{0.154} = 12.83$$

Exemplo 2: Análise pelos valores críticos

Passo 3: Calculamos os valores críticos na distribuição t de Student:



Passo 4: Decidimos se aceitamos ou rejeitamos a hipótese nula:

Como t = 12.83 caiu na região crítica, rejeitamos a hipótese nula (de que não há nenhuma correlação linear). Logo, há uma correlação linear entre as variáveis. Isto é, há evidência suficiente para apoiar a afirmativa da existência de uma correlação linear entre as variáveis.

Observação

Cuidado: A existência de correlação entre duas variáveis não implica que uma causa a outra.

Por exemplo, no exercício anterior podemos concluir que há uma correlação entre os custos de pizza e as tarifas do metrô, mas não podemos concluir que aumentos no custo da pizza cause aumento nas tarifas de metrô.

Exercício 1

O alongamento (y) de uma mola foi medido em função da carga (x) aplicada. Os resultados estão listadas na tabela a seguir.

Carga (kg)	Alongamento (cm)
4	7.3
5	8.5
6	9.0
7	9.5
8	9.9

a) Calcular o coeficiente de correlação linear.
b) Construir um diagrama de dispersão (x,y).
c) Testar se a correlação é significativa, ao nível de significância de 5%.

Regressão

Regressão

Se existir uma correlação linear entre **duas variáveis**, a análise de **regressão** serve para **encontrar** a **função linear que melhor se ajusta aos dados**.

Equação de regressão

Dada uma coleção de dados amostrais emparelhados (x, y), a **equação de regressão**

$$\hat{y} = \underbrace{a}_{\text{coeficiente angular}}x + \underbrace{b}_{\text{coeficiente linear}}$$

descreve algebricamente a relação entre as variáveis x e y.

O **gráfico da equação de regressão** é chamado de **reta de regressão** (**reta de melhor ajuste**, ou **reta de mínimos quadrados**).

Objetivo

Achar a equação da reta de regressão.

Notação

$n:$ *tamanho da amostra de dados emparelhados* $(x\ ,\ y)$

$r:$ *coeficiente de correlação linear amostral*

$a:$ *coeficiente angular da reta*

$b:$ *coeficiente linear da reta*

$\bar{x}$ *e* $\bar{y}:$ *médias amostrais das variáveis* x *e* y

s_x *e* $s_y:$ *desvios padrões amostrais das variáveis* x *e* y

Requisitos

1) A amostra é uma amostra aleatória simples.
2) O gráfico de dispersão (x,y) se aproxima de uma reta.
3) Valores atípicos devem ser removidos caso se saiba que são erros.

Fórmulas para encontrar os coeficientes a e b

Os **coeficientes angular a** e **linear b** da **reta de regressão** são dados por:

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} \qquad b = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

Ou:

$$a = r\frac{s_y}{s_x} \qquad b = \bar{y} - a\bar{x}$$

Exemplo 3

Determine a **equação de regressão para a amostra de dados** que relaciona **custos de pizza (x) com tarifas de metrô (y)**.

Cálculo dos coeficientes:

$$a = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{6(9.4575) - (6.50)(6.35)}{6(9.77) - (6.50)^2} = 0.945$$

$$b = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{(6.35)(9.77) - (6.50)(9.4575)}{6(9.77) - (6.50)^2} = 0.0346$$

Equação da reta de regressão:

$$\hat{y} = ax + b$$

$$\hat{y} = 0.945x + 0.0346$$

Exemplo 4: predições

Use a **equação de regressão para predizer** o **custo da tarifa de metrô** quando uma fatia de **pizza custar 2.50 dólares**.

$$\hat{y} = 0.945x + 0.0346$$

$$\hat{y} = 0.945(2.50) + 0.0346$$

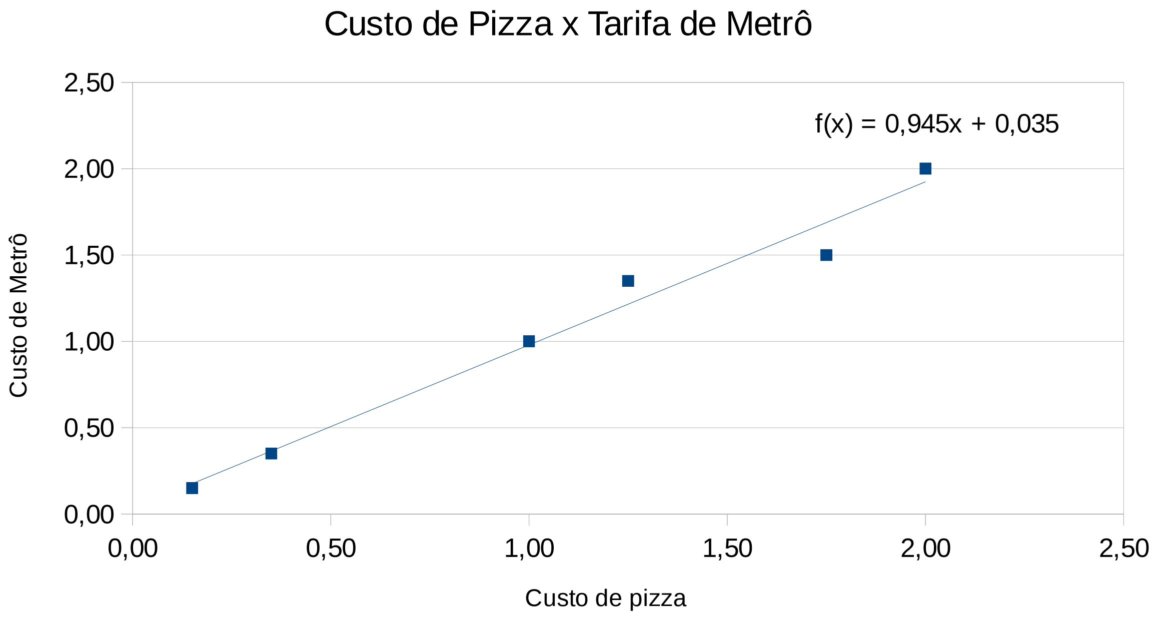
$$\hat{y} = 2.40$$

Logo, a previsão é que a tarifa de metrô passe a custar $ 2.40 quando a pizza custar $ 2.50.

Obs: tome cuidado ao predizer valores de y para valores de x que estão muito distantes dos dados amostrais. Pois tais previsões podem resultar em grandes erros.

Exemplo 5: plotando a reta de regressão

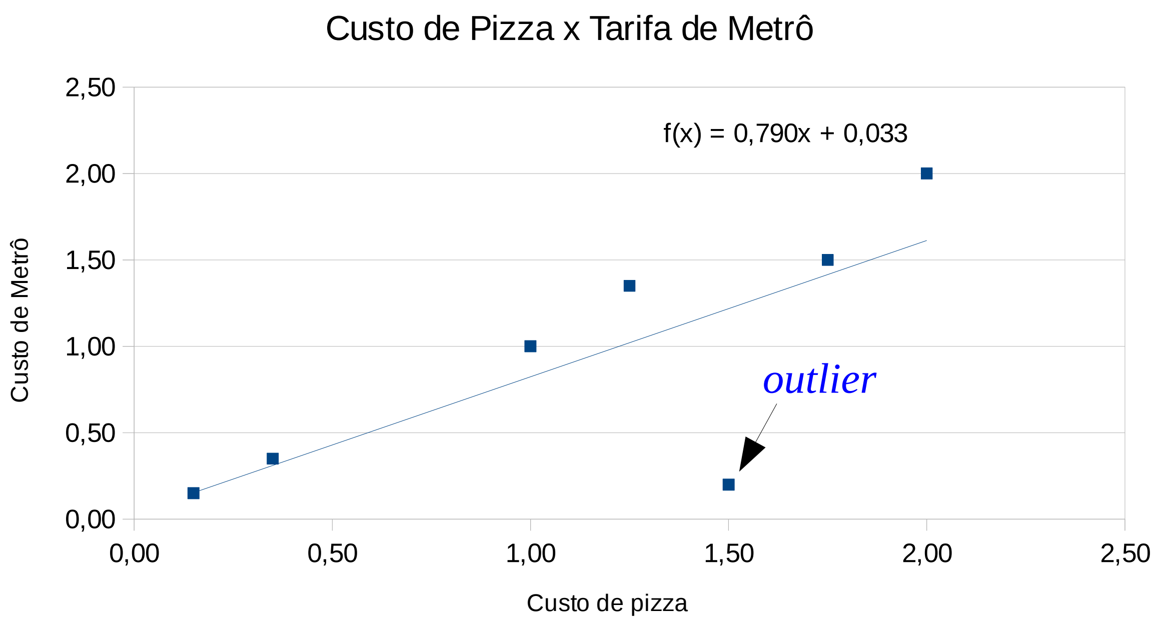
Plote a reta de regressão do exemplo anterior **usando uma planilha eletrônica**.



Obs: ver arquivo “regressao.xls”. Em planilhas eletrônicas, retas de regressão são geralmente denominadas por linhas de tendência.

Exemplo 6: valores atípicos (ou outliers)

O gráfico de dispersão abaixo ilustra um conjunto de dados com valores atípicos ou outliers.

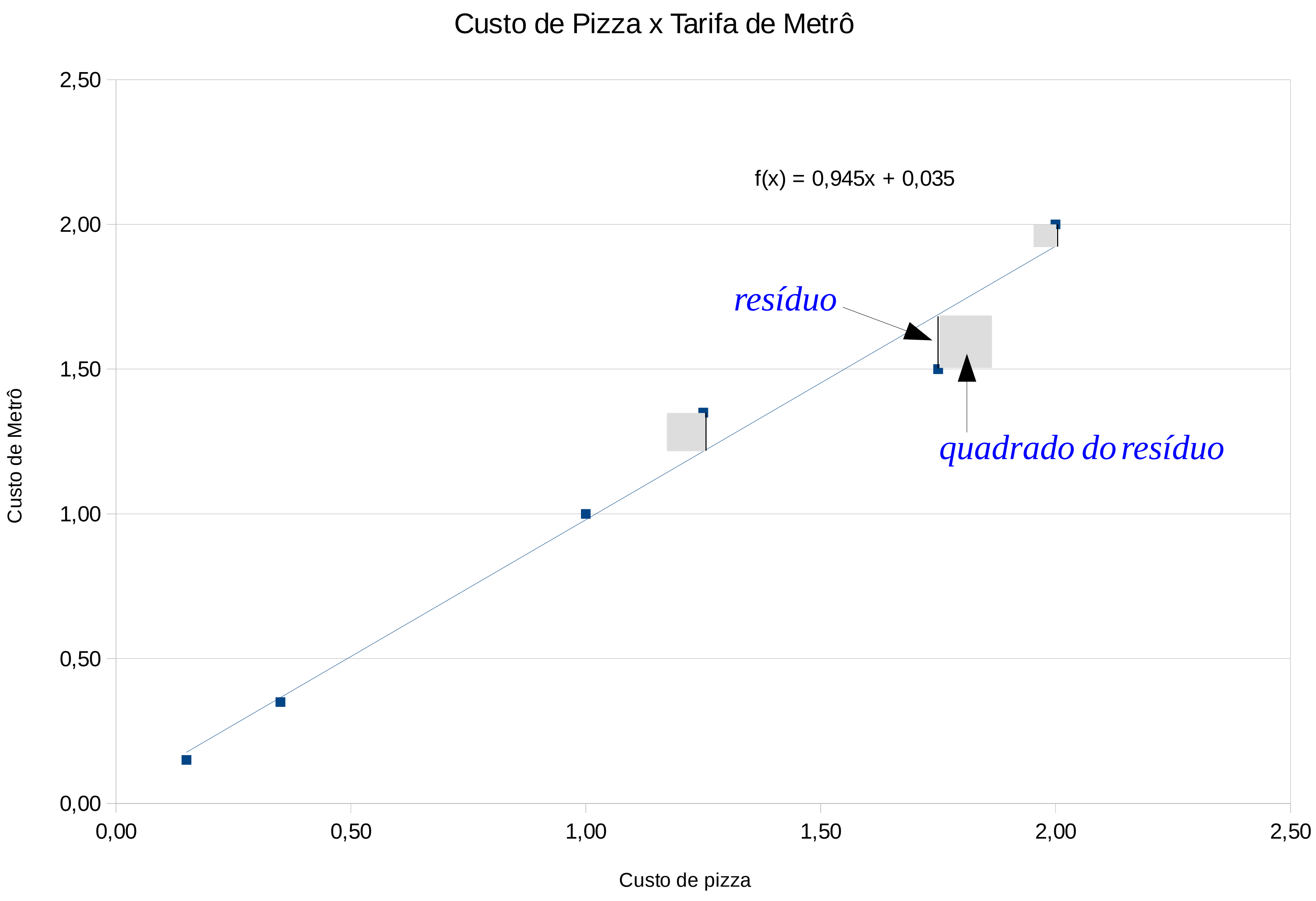


Resíduos e propriedade dos mínimos quadrados

Resíduo: é a diferença entre um valor amostral observado y e seu valor correspondente predito ŷ.

Mínimos quadrados: uma reta de regressão satisfaz a propriedade dos mínimos quadrados se a soma dos quadrados dos resíduos for a menor possível.

Propriedade dos mínimos quadrados



Exercício 2: experimento da mola

Colete dados do alongamento (y) de uma mola quando esta é submetida a uma carga (x). Use a planilha eletrônica para:

Carga (g)	Alongamento (cm)

a) Construir um diagrama de dispersão (x,y).
b) Calcular os coeficientes de correlação linear e os coeficientes angular e linear da reta de regressão.
c) Ajustar a reta de regressão.
d) Calcular os resíduos de cada ponto de dado.