Conceitos básicos

Variáveis

- As **características medidas são conhecidas como variáveis**. Por exemplo:
  - Estudo sobre habitantes de uma cidade, as variáveis podem ser: Altura, sexo, cor do cabelo, cor dos olhos, etc

- Divididas em dois tipos:
  - **Independente**
  - **Dependente**

Tipos de variáveis

Independente:

- Valores manipulados ou selecionados pelo pesquisador (meio, idade, mês).

Dependente:

- Valores observados, contados, medidos, ... que não estejam sob controle direto do pesquisador (velocidade, taxa de câmbio).

Tipos de variáveis

- **A variável 'dependente' é esta que analisamos em função dos valores de uma outra variável.**

Correlação e regressão

Correlação e regressão

- As técnicas de **correlação e regressão** analisam dados amostrais, **procurando determinar como duas (ou mais) variáveis estão relacionadas umas com as outras**.

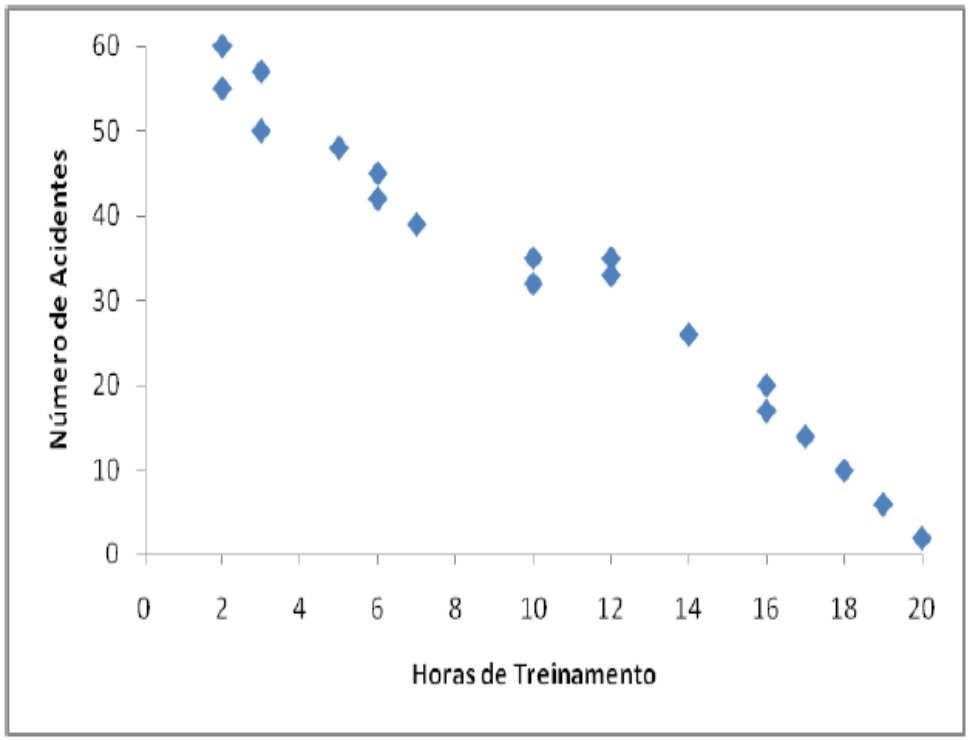
Variável Independente	Variável Dependente
Horas de treinamento	Número de acidentes
Número do sapato	Altura da pessoa
Cigarros por dia	Capacidade pulmonar
Meses do ano	Volume de vendas
Peso da pessoa	QI

Correlação e regressão

- A **análise de correlação** tem como resultado um **número que expressa o grau de relacionamento** entre duas variáveis.

- A **análise de regressão** expressa o resultado em uma **equação matemática**, descrevendo o relacionamento.

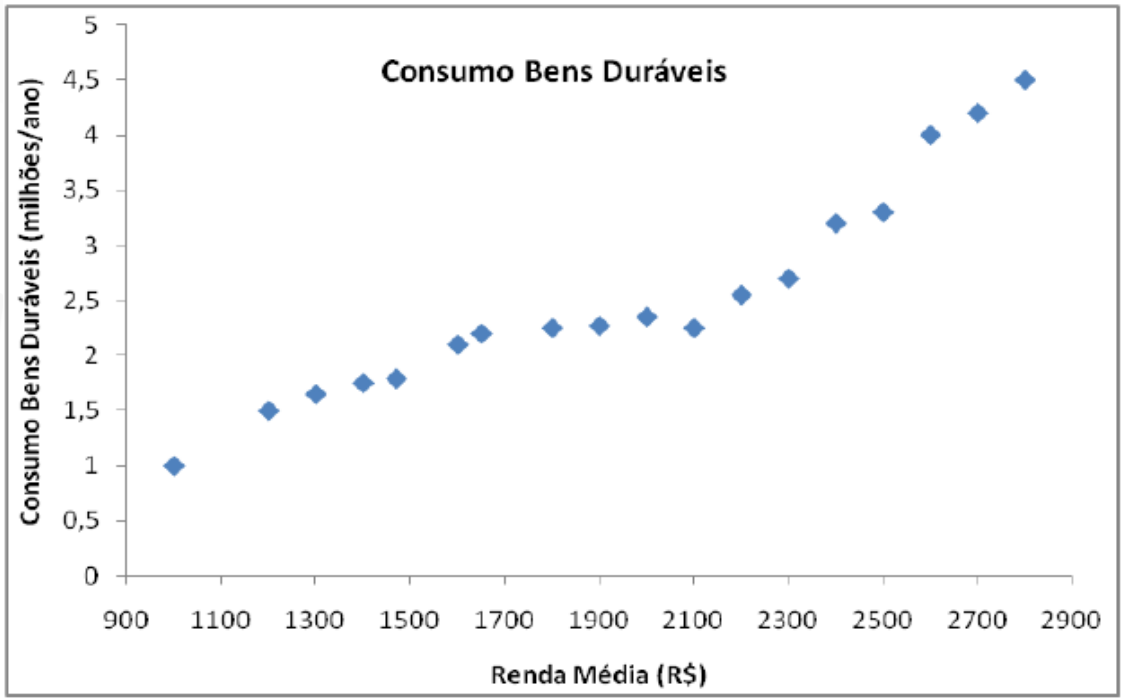
Correlação



A análise gráfica do comportamento entre as variáveis mostra a **existência de correlação negativa**, pois à medida que X cresce, Y decresce

O gráfico mostra que a empresa, ao investir em treinamento, reduz o número de acidentes na fábrica

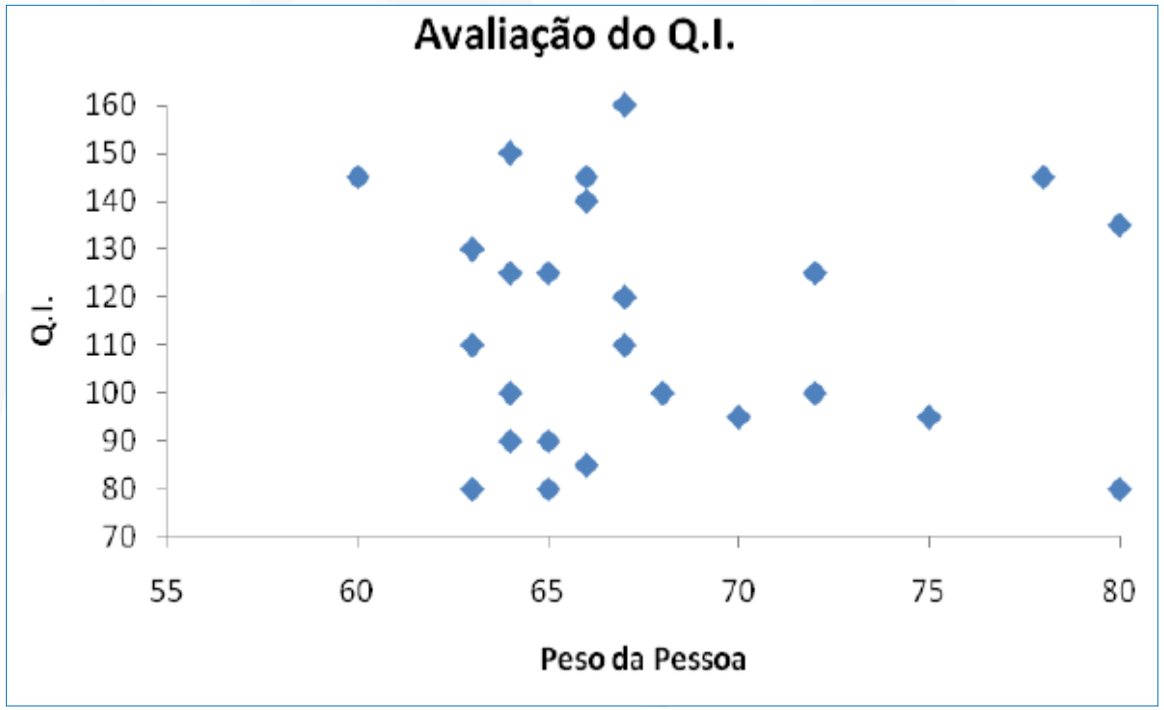
Correlação



A análise gráfica do comportamento entre as variáveis mostra a **existência de correlação positiva**, pois à medida que X cresce, Y também cresce.

O gráfico mostra que, com o aumento médio da renda da população, o consumo de bens duráveis aumenta.

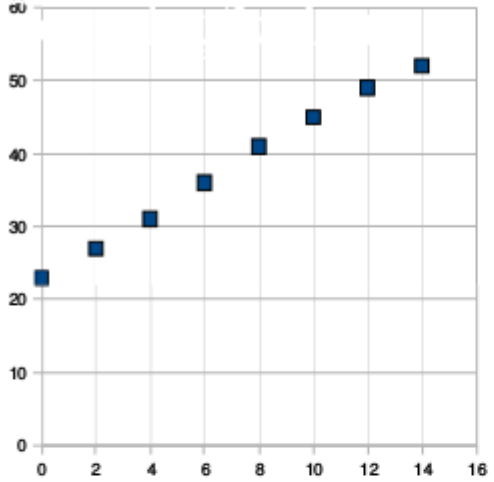
Correlação



Não há correlação linear, o gráfico mostra que **não existe evidência de alguma relação** entre o peso de uma pessoa com seu Q.I.

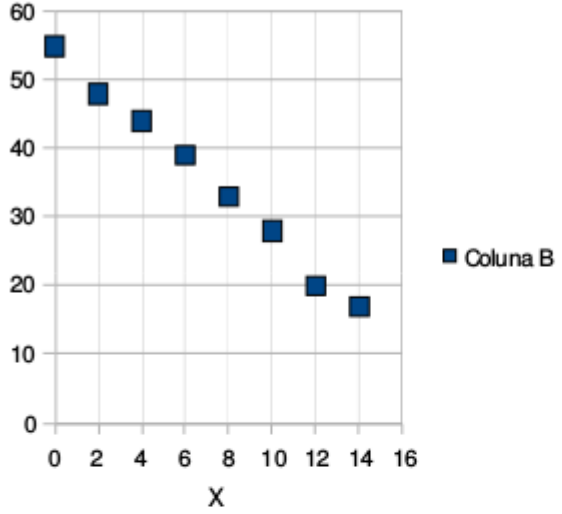
Exemplos

Correlação Linear Positiva



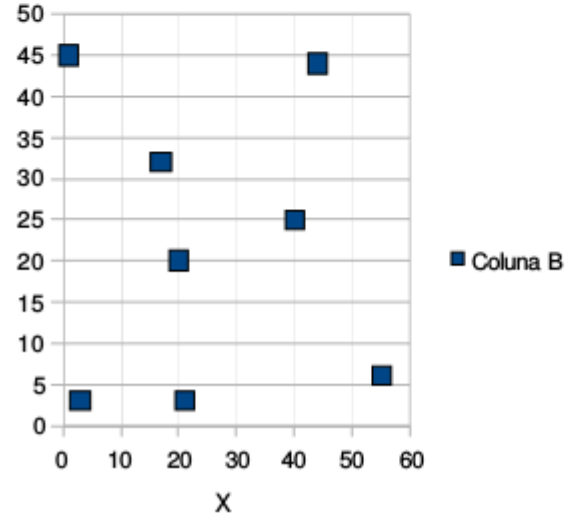
À medida que x cresce, y tende a crescer.

Correlação Linear Negativa

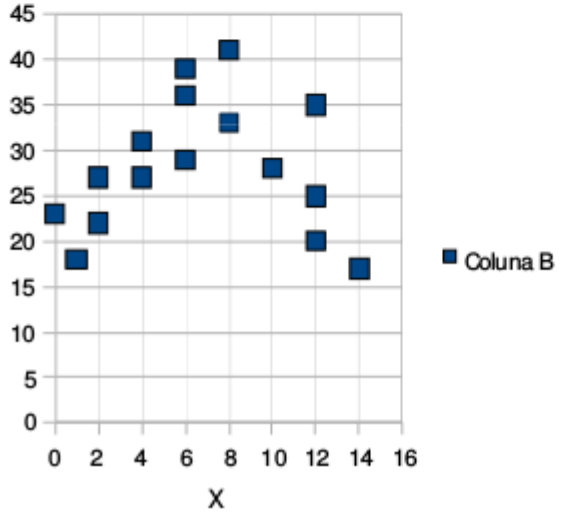


À medida que x cresce, y tende a decrescer.

Não há Correlação



Correlação Não Linear



Correlação: tipos

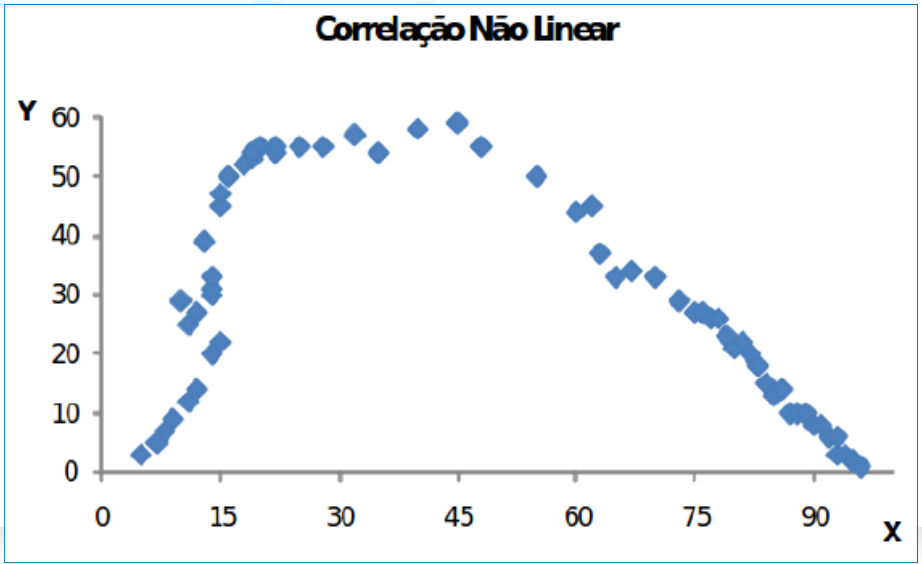
Podemos ter dois tipos de correlação entre as variáveis:

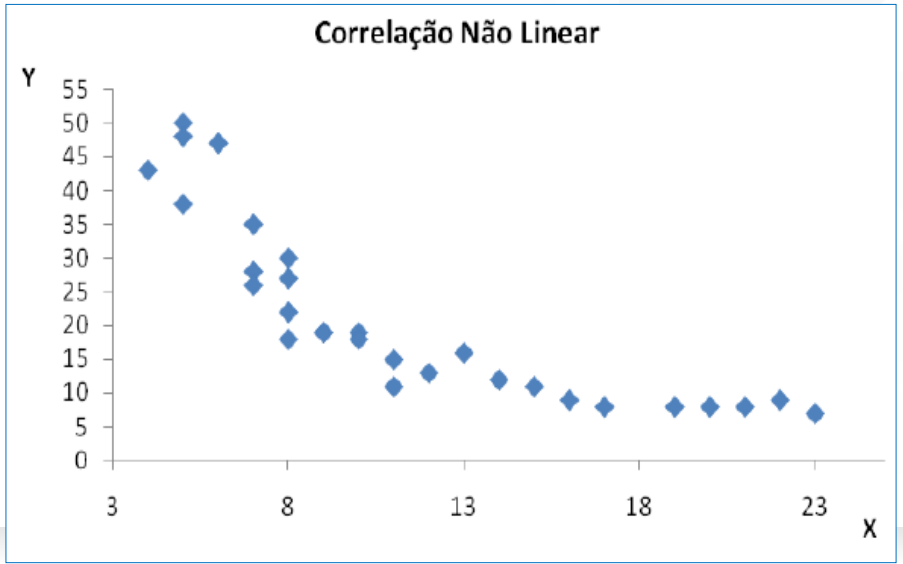
- **Correlação linear**

  em que a relação entre as duas variáveis é expressa adequadamente por uma reta.

- **Correlação não-linear**

  Apesar de existir uma relação clara entre as variáveis, esta não pode ser modelada por uma reta.





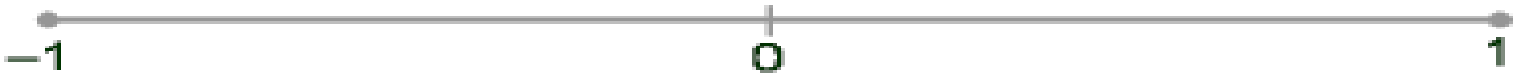
Coeficiente de correlação

Coeficiente de correlação

- Utilizar apenas o **gráfico de dispersão** para interpretar a existência de uma correlação **pode ser uma tarefa bastante subjetiva**.
- Como **medida mais objetiva**, utiliza-se medir o grau e o tipo de uma correlação linear entre duas variáveis por meio do cálculo do **coeficiente de correlação**.

Coeficiente de correlação

O intervalo de variação do **coeficiente de correlação r** está ente -1 à 1.



Coeficiente de correlação

$$r_{XY} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}}$$

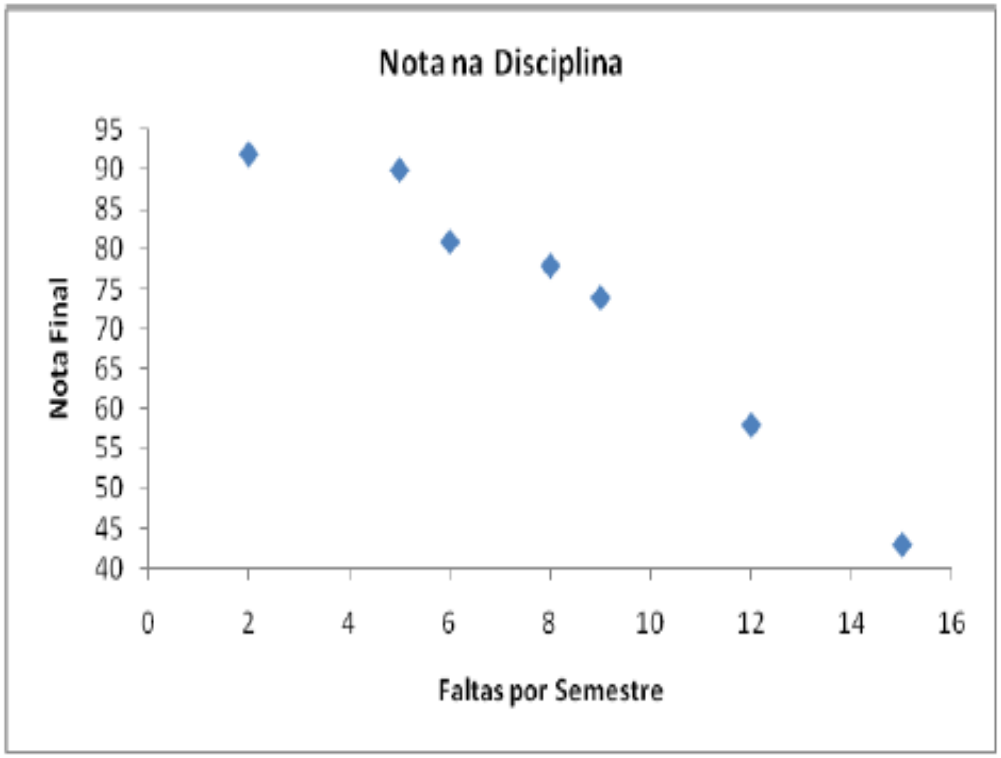
Coeficiente de correlação

Como exemplo, analisaremos o **coef. de correlação entre o número de faltas dos alunos por semestre, em relação a suas respectivas notas finais** em uma determinada disciplina.

Faltas por semestre (X)	Nota Final (Y)
8	78
2	92
5	90
12	58
15	43
9	74
6	81

$$r_{XY} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}}$$

r = -0.975



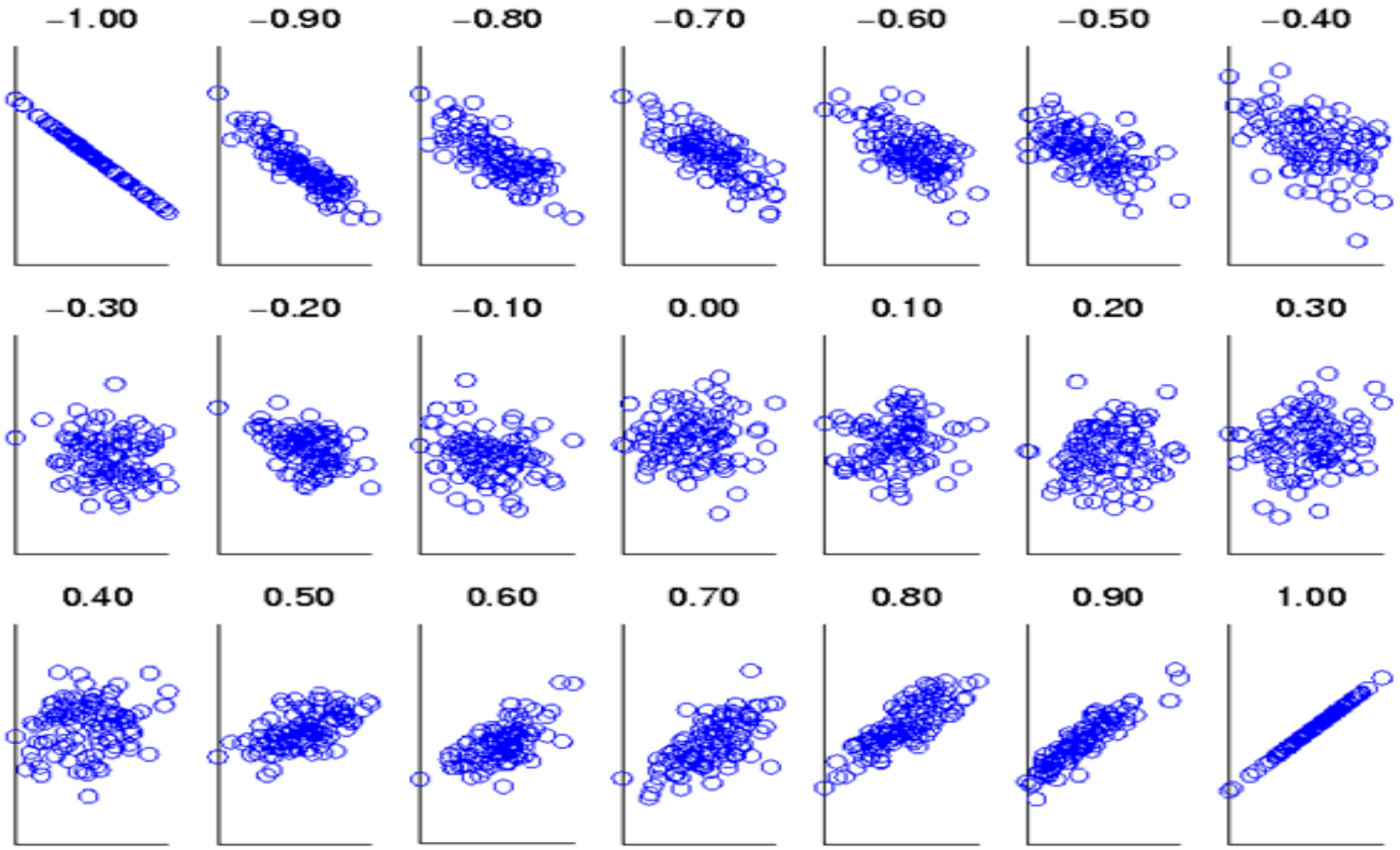
Coeficiente de correlação

Interpretação do coeficiente de correlação

Valor de ρ (+ ou −)	Interpretação
0.00 a 0.19	Uma correlação bem fraca
0.20 a 0.39	Uma correlação fraca
0.40 a 0.69	Uma correlação moderada
0.70 a 0.89	Uma correlação forte
0.90 a 1.00	Uma correlação muito forte

http://leg.ufpr.br/~silvia/CE003/node74.html

Coeficiente de correlação



Coeficiente de determinação

Coeficiente de determinação

- O quadrado do coeficiente de correlação (de Pearson) é chamado de **coeficiente de determinação ($r^2=[0,1]$)**.

- **É uma medida da proporção da variabilidade em uma variável que é explicada pela variabilidade da outra.**

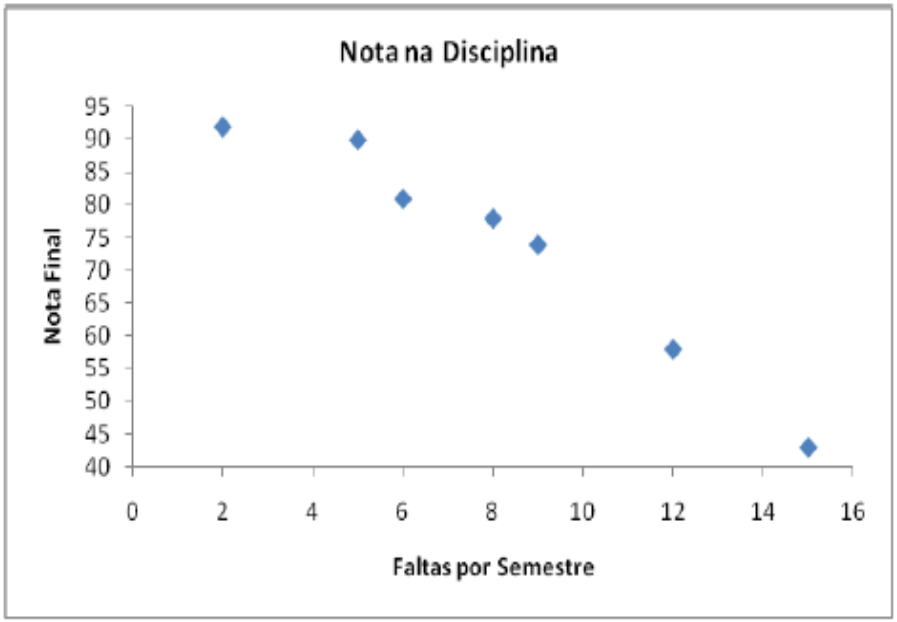
  - Na prática, é pouco comum que tenhamos uma correlação perfeita $r^2=1$

    pois existem muitos fatores que determinam as relações entre variáveis na vida real.

Coeficiente de determinação

Relação entre o número de faltas dos alunos por semestre, e suas notas finais.

Faltas por semestre (X)	Nota Final (Y)
8	78
2	92
5	90
12	58
15	43
9	74
6	81



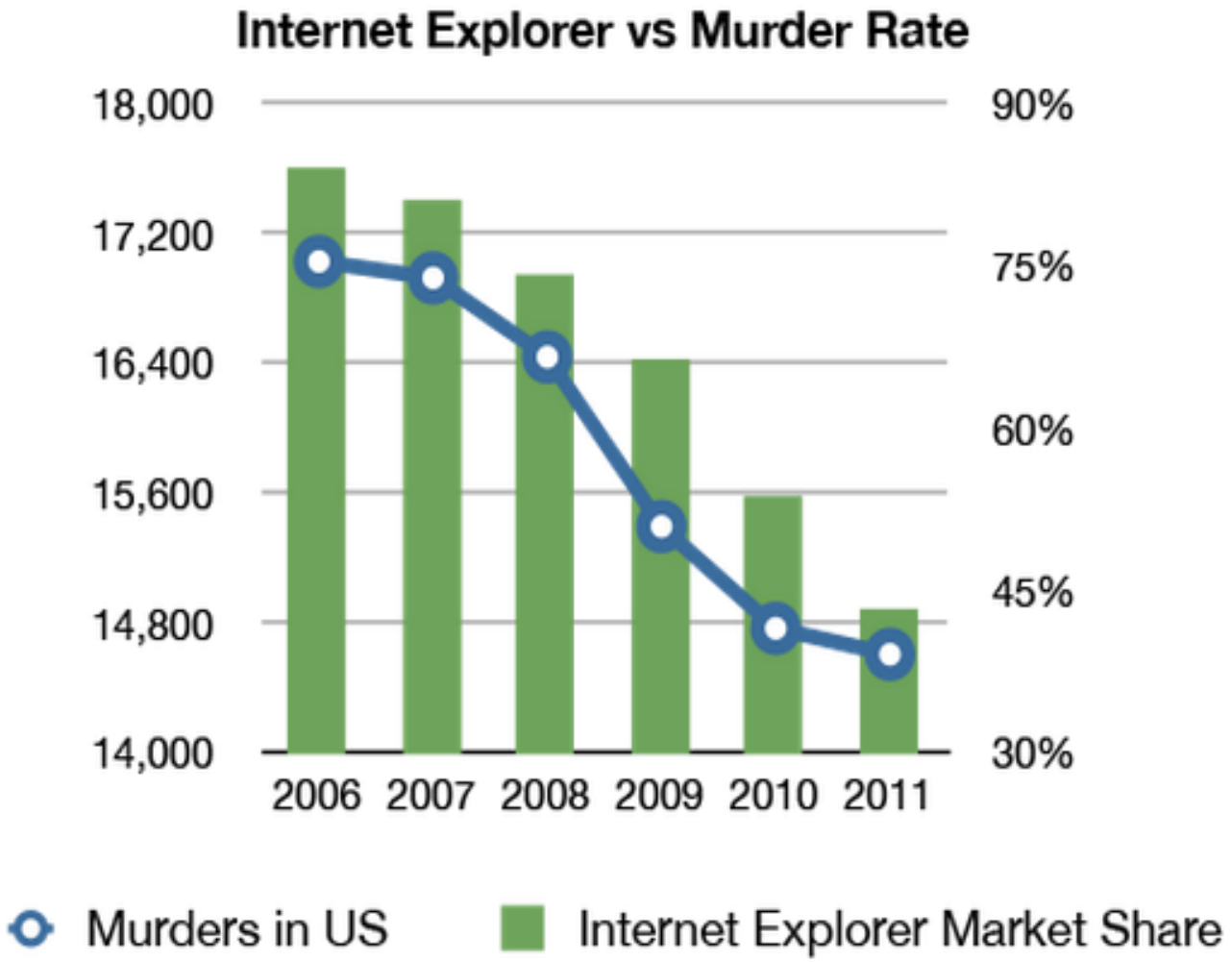
$$r = \frac{7(3.751) - (57)(516)}{\sqrt{7(579) - (57)^2}\sqrt{7(39.898) - (516)^2}}$$

$$r = -0.975$$

r² = 0.9501633 (ou 95%)

Então cerca de 5% da variabilidade da nota final ***não pode ser descrito ou explicado*** *pela variabilidade do número de faltas por semestre e vice-versa.*

Causalidade?



Causalidade e correlação

Correlação não necessariamente implica em causalidade.

- Pesquisadores frequentemente são tentados a inferir **uma relação de causa e efeito entre X e Y**, quando eles ajustam um modelo de regressão, ou realizam uma análise de correlação

- **Uma associação significativa entre X e Y não necessariamente implica em uma relação de causa e efeito**

Curvas de regressão

Reta de regressão linear

Depois de constatar que existe uma **correlação linear significante**, é possível escrever uma **equação que descreva a relação linear** entre as variáveis X e Y.

Essa equação chama-se reta de regressão, ou **reta do ajuste ótimo**

Pode-se escrever a equação de uma reta como ***y* = *mx* + *b***, onde ***m*** é ***a*** inclinação da reta e ***b***, o intercepto ***y***. Assim, a reta de regressão é:

$$\hat{Y} = mX + b$$

A inclinação m é dada por:

$$m = \frac{N\sum_{i=1}^{N} XY - \sum_{i=1}^{N} X \sum_{i=1}^{N} Y}{N\sum_{i=1}^{N} X^2 - (\sum_{i=1}^{N} X)^2}$$

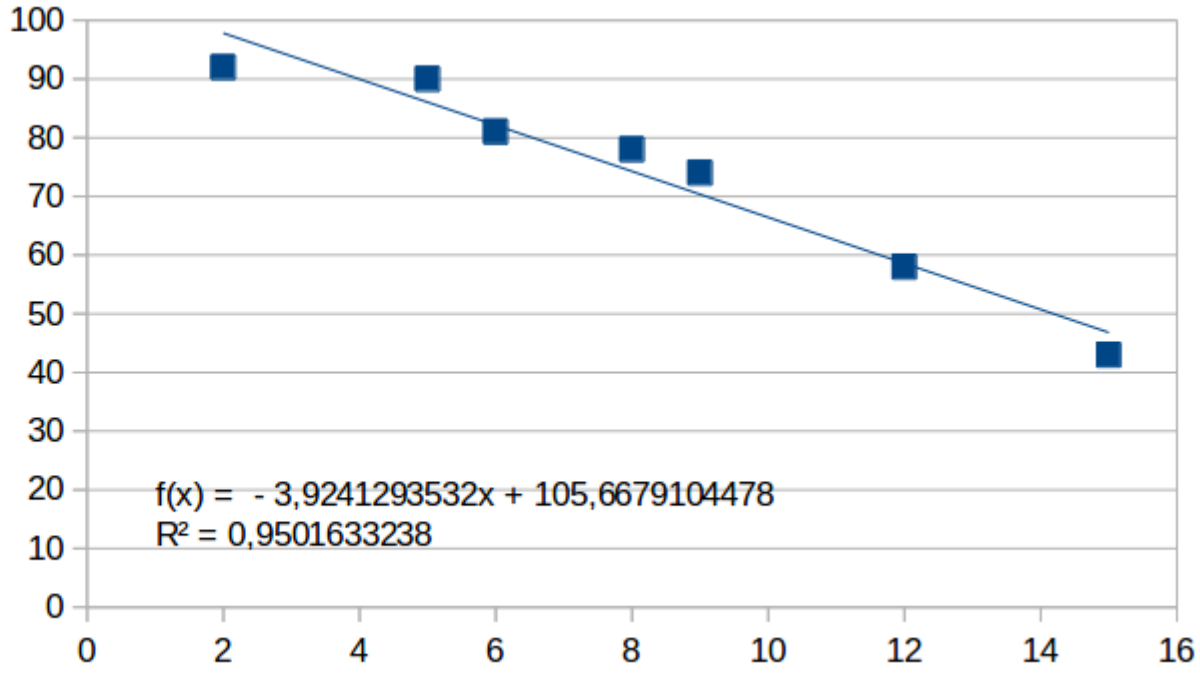
E o intercepto y é:

$$b = \overline{Y} - m\overline{X}$$

Curvas de regressão

8	78
2	92
5	90
12	58
15	43
9	74
6	81

r = -0,974763214



Exercício

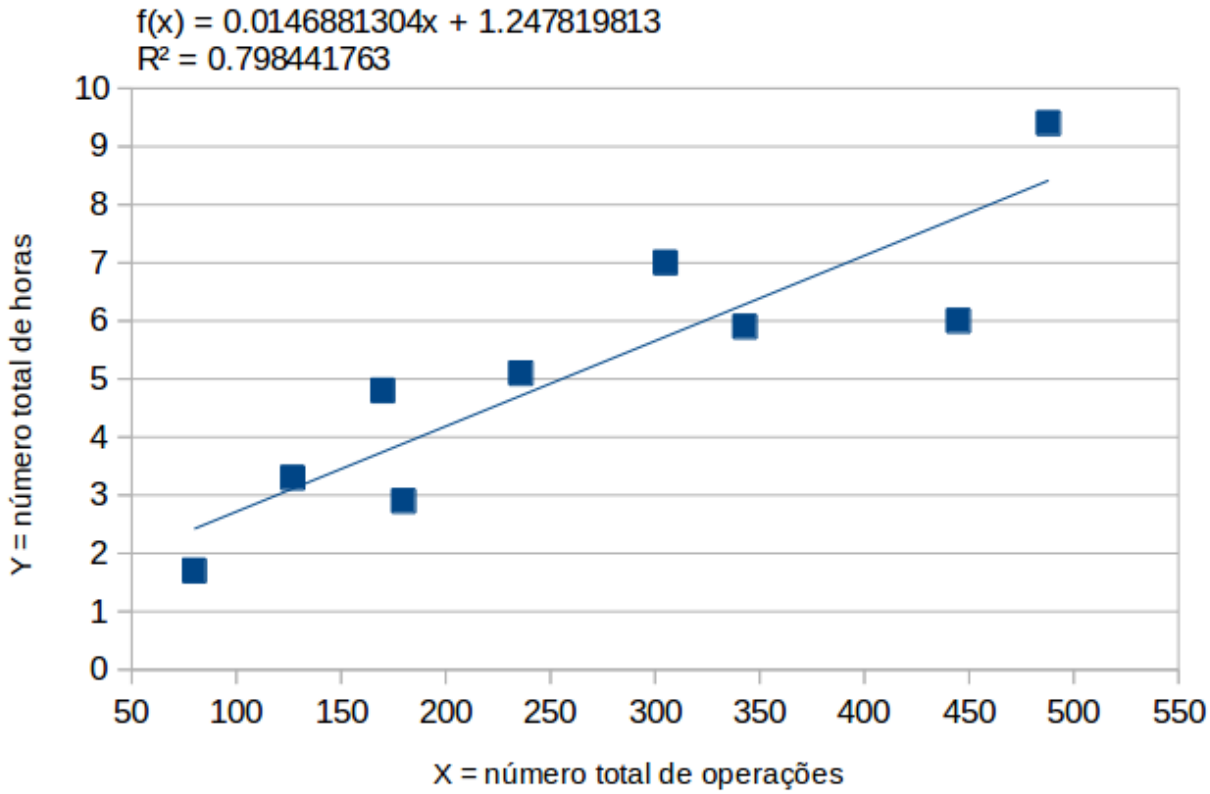
Para estudar a relação entre:
- X (número total de operações de furar e rebitar), e
- Y (número total de horas necessárias à montagem da parte de uma estrutura),
Registraram-se os dados da tabela abaixo.

estudo	A	B	C	D	E	F	G	H	I
X	236	80	127	445	180	343	305	488	170
Y	5,1	1,7	3,3	6,0	2,9	5,9	7,0	9,4	4,8

- Faça o gráfico destes dados com número total de operações no eixo x.
- Calcule o coeficiente de correlação linear (r) para estes dados e cheque se o valor obtido parece consistente com seu gráfico.
- Qual proporção da variabilidade no total de operações de furar e rebitar que pode ser explicada pelo número total de horas necessárias à montagem da parte de uma estrutura?

Exercício

X	Y
236	5.1
80	1.7
127	3.3
445	6
180	2.9
343	5.9
305	7
488	9.4
170	4.8



Coeficiente de correlação = 0.8936
Coeficiente de determinação = 0.7984