# Correlação e regressão (revisitado)

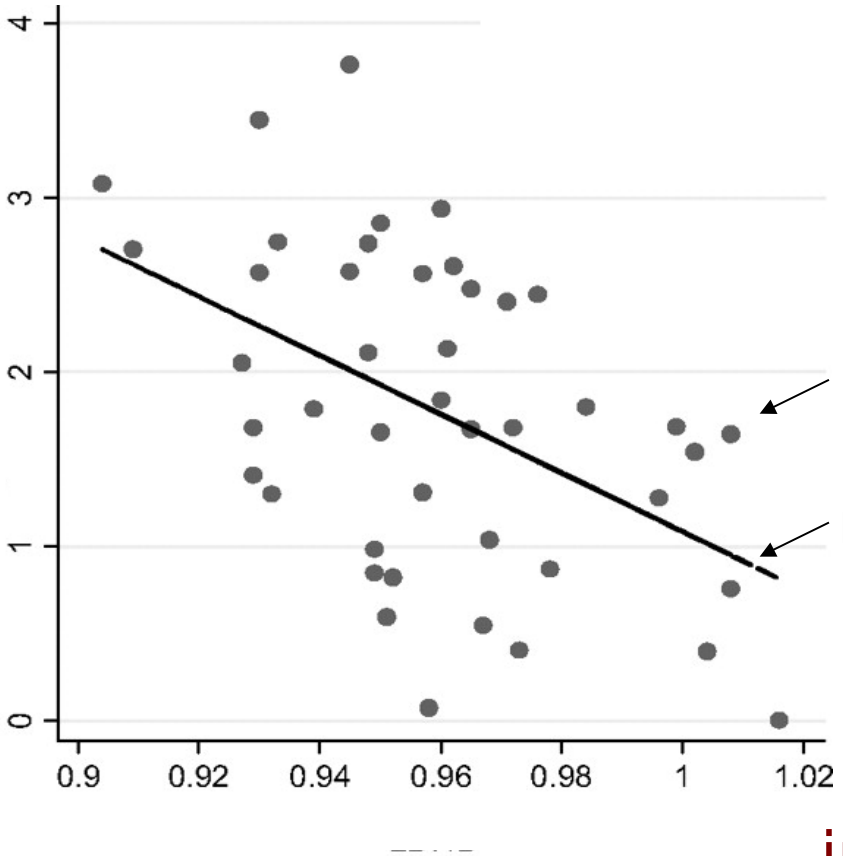

- Ronaldo C. Prati
  ronaldo.prati@ufabc.edu.br

# Relembrando o final a última aula...

# Gráficos de dispersão

Índice de
sucesso financeiro

dependente:
contínua




Razão indicador/anular

independente:
contínua

John M. Coates, Mark Gurnell, and Aldo Rustichini
Second-to-fourth digit ratio predicts success among high-frequency financial traders
PNAS 2009 106:623-628; published online before print January 12, 2009, doi:10.1073/pnas.0810907106

# 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

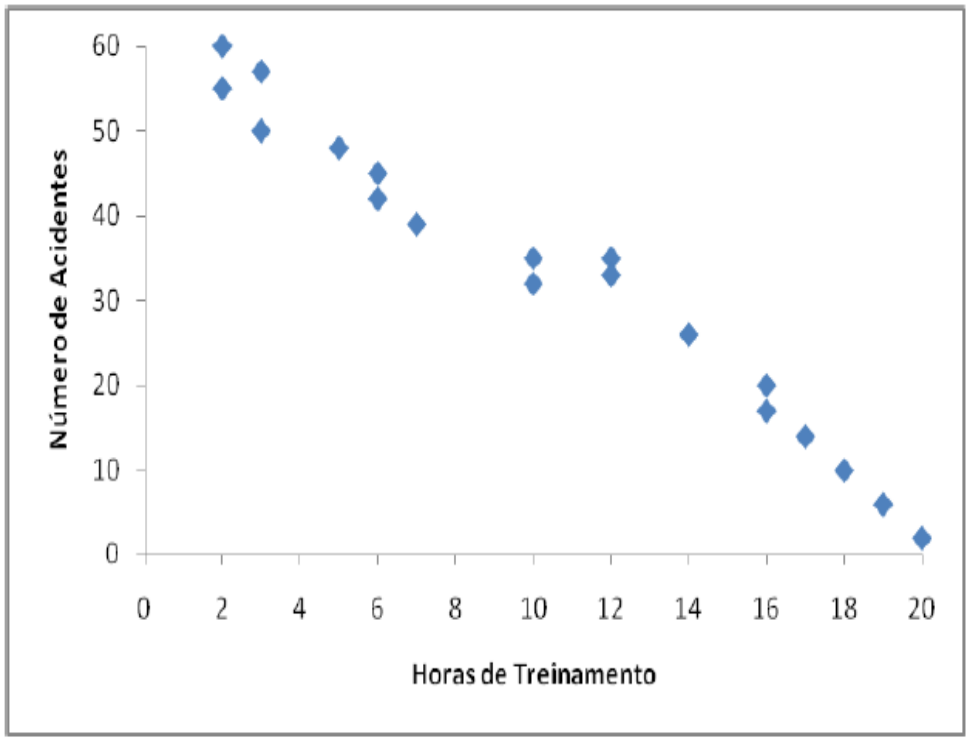
- A **análise de correlação** tem como resultado um **número que expressa o grau de relacionamento** entre duas variáveis.

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

Ambas análises, geralmente utilizadas em pesquisas exploratórias.

# Correlação

Variável dependente

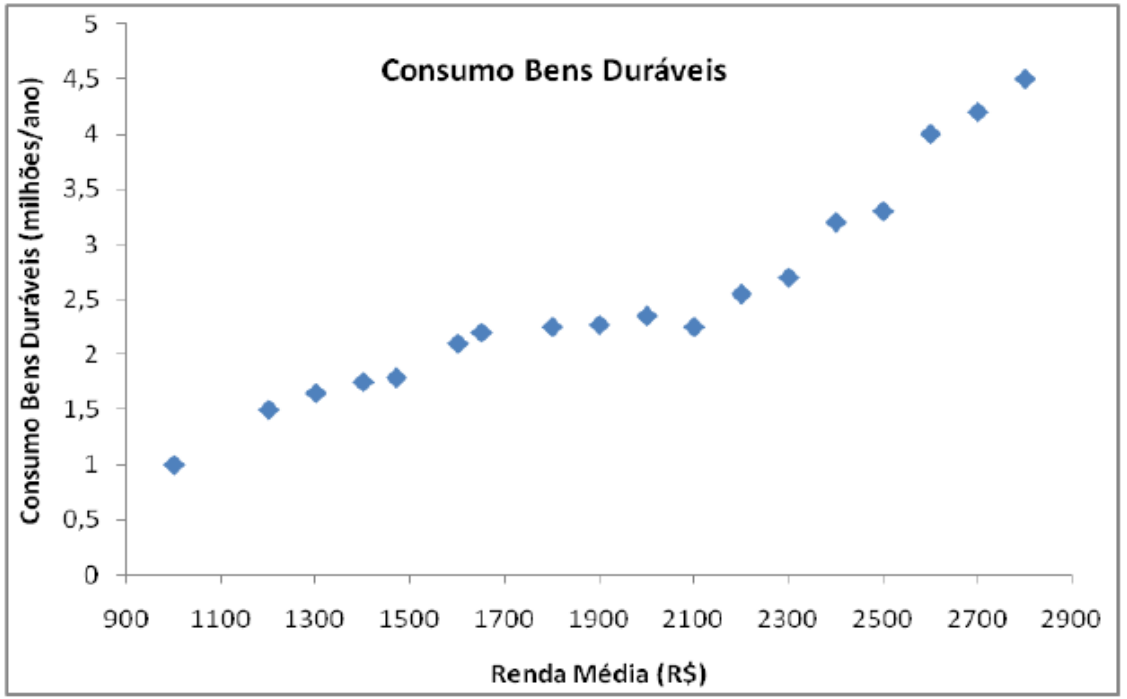



Variável independente

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

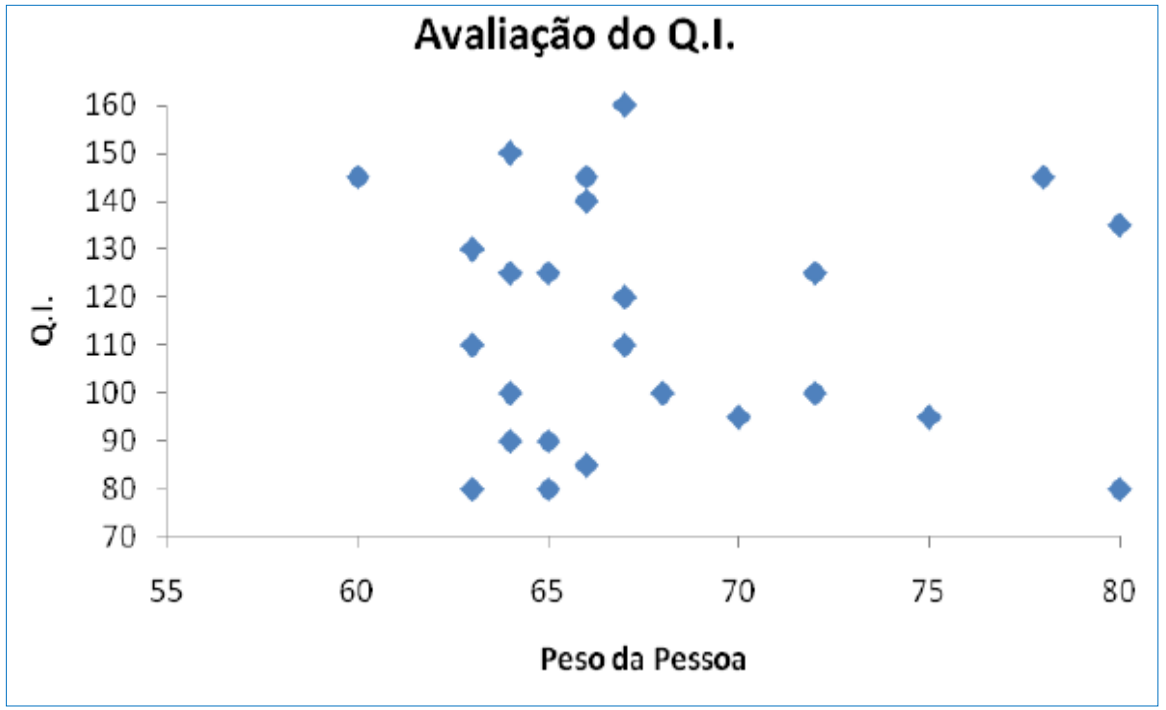
# Correlação



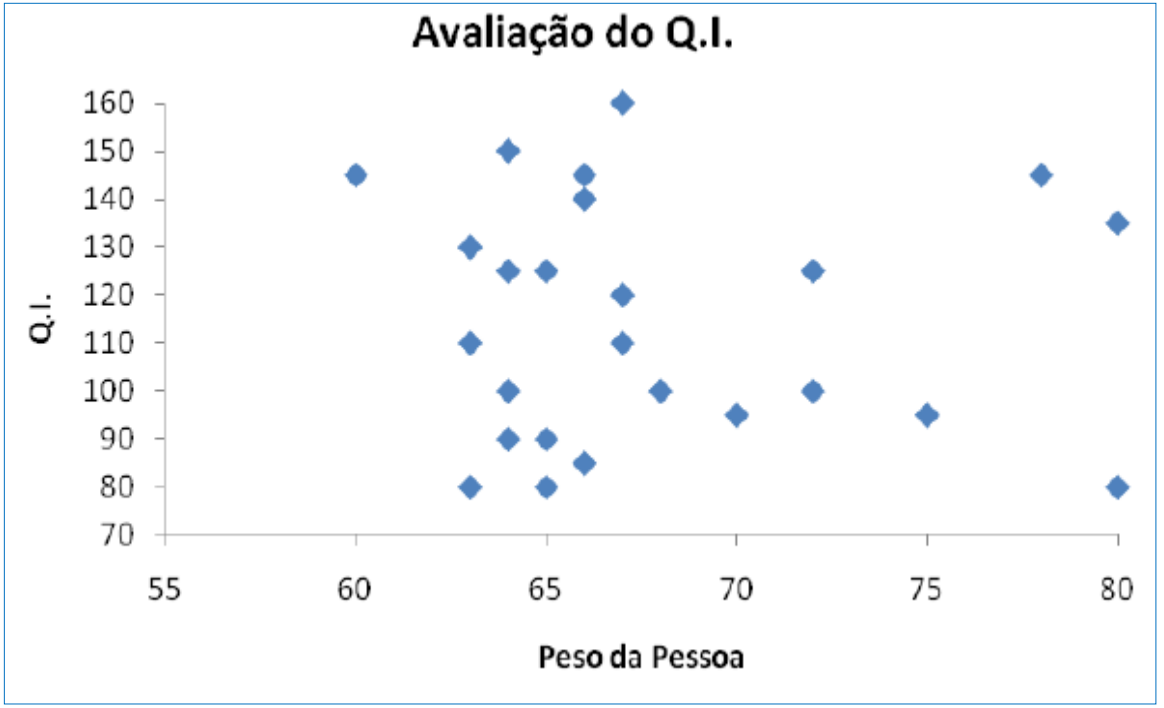

A análise gráfica do comportamento entre as variáveis mostra a **existência de correlação positiva**, pois à medida que X cresce, Y também cresce.

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

# Correlação


Avaliação do Q.I.
Q.I.
160
150
140
130
120
110
100
90
80
70
55
60
65
70
75
80
Peso da Pessoa

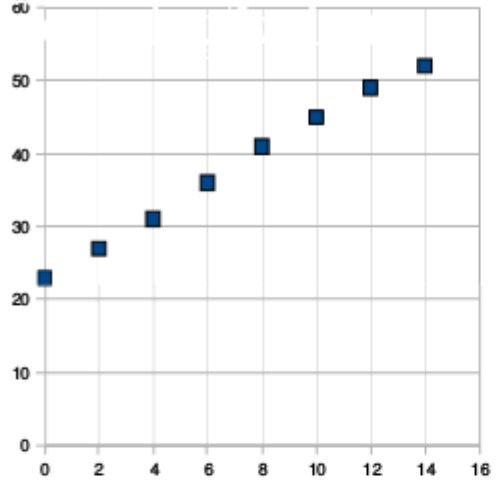
# Correlação




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

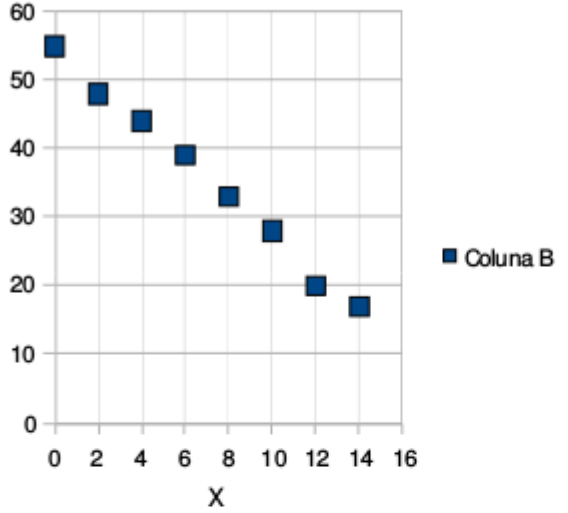
# Exemplos

## Correlação Linear Positiva



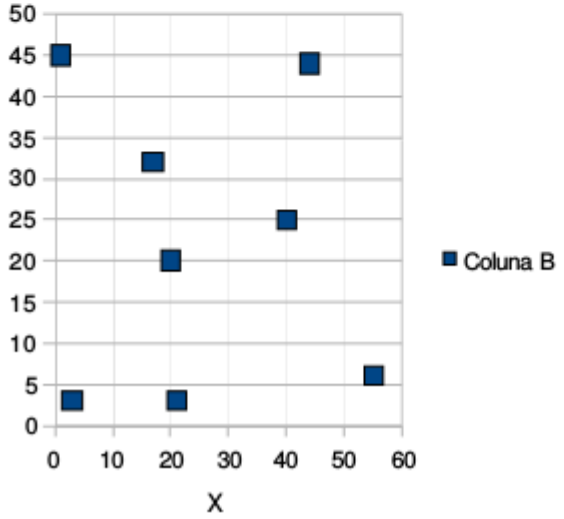

À medida que x cresce, y tende a crescer.

## Correlação Linear Negativa

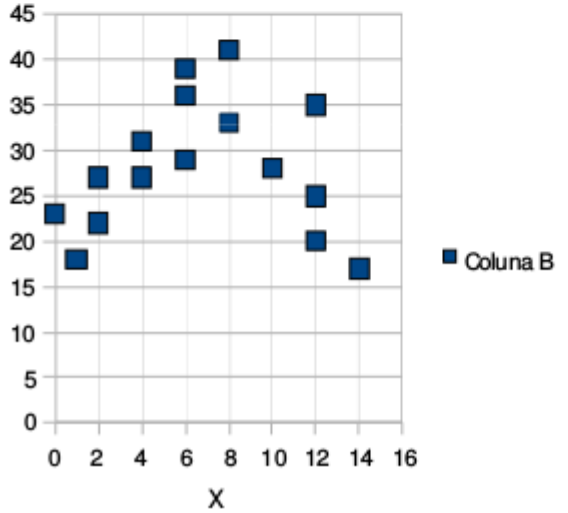



À medida que x cresce, y tende a decrescer.

## Não há Correlação




## Correlação Não Linear



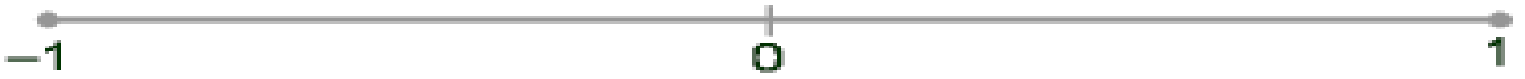
# Coeficiente de correlação

# 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.




**Valor de r próximo de -1:** as variáveis X e Y têm forte correlação linear negativa

**Valor de r próximo de zero:** se não existir, ou se existir pouca correlação linear entre as variáveis X e Y

**Valor de r próximo de 1**: as variáveis X e Y têm forte correlação linear positiva

# Coeficiente de correlação

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

# Coeficiente de correlação

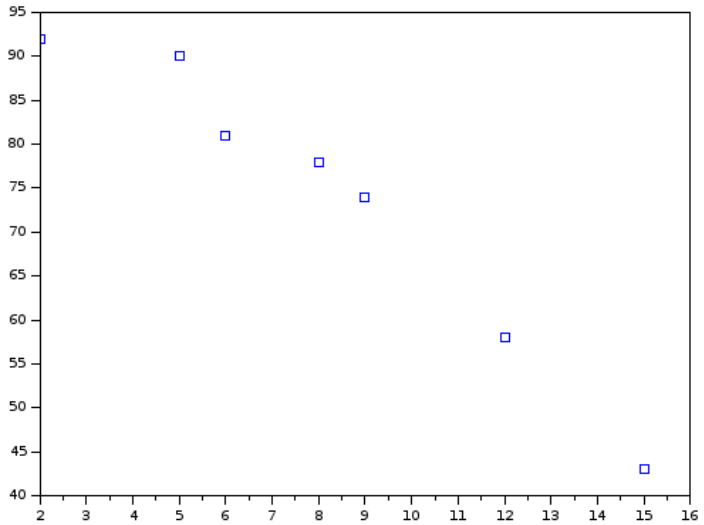
$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\operatorname{cov}(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}(x_i - x_j)\cdot(y_i - y_j) = \frac{1}{n^2}\sum_{i}\sum_{j>i}(x_i - x_j)\cdot(y_i - y_j).$$

-->x = [8, 2, 5, 12, 15, 9, 6];

-->y = [78, 92, 90, 58, 43, 74, 81];

-->plot(x,y,'s');




-->correl(x,y)

ans =

- 0.9747632

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

-->N = length(x);

-->**num** = N*sum(x.*y) - sum(x)*sum(y);

-->**den1** = sqrt(N*sum(x.^2)-sum(x)^2);

-->**den2** = sqrt(N*sum(y.^2)-sum(y)^2);

-->r = **num**/(**den1***__den2__)

- 0.9747632

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

# 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.**

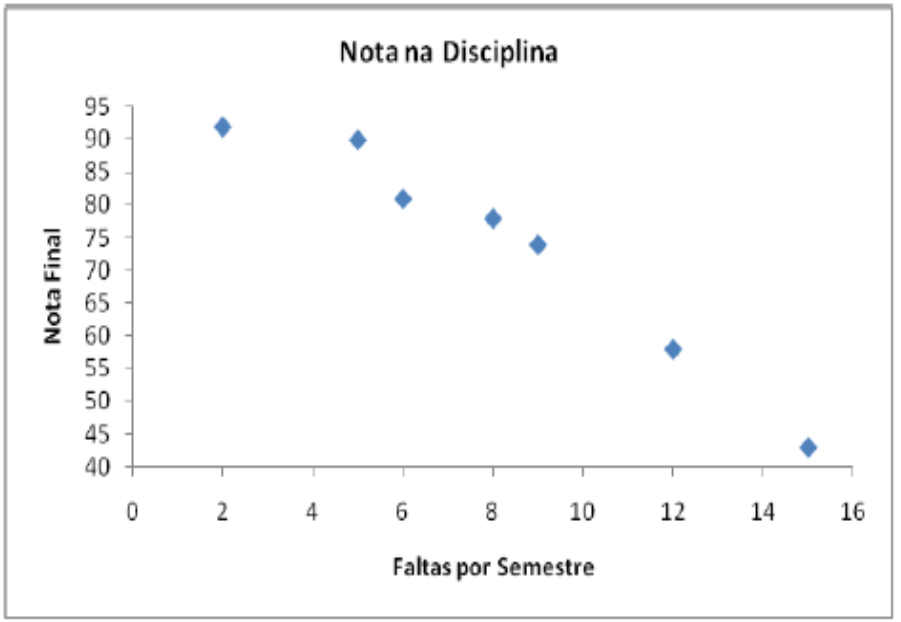
  - Na prática, é pouco comum que tenhamos uma correlação perfeita $r^2=1$

    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 |




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

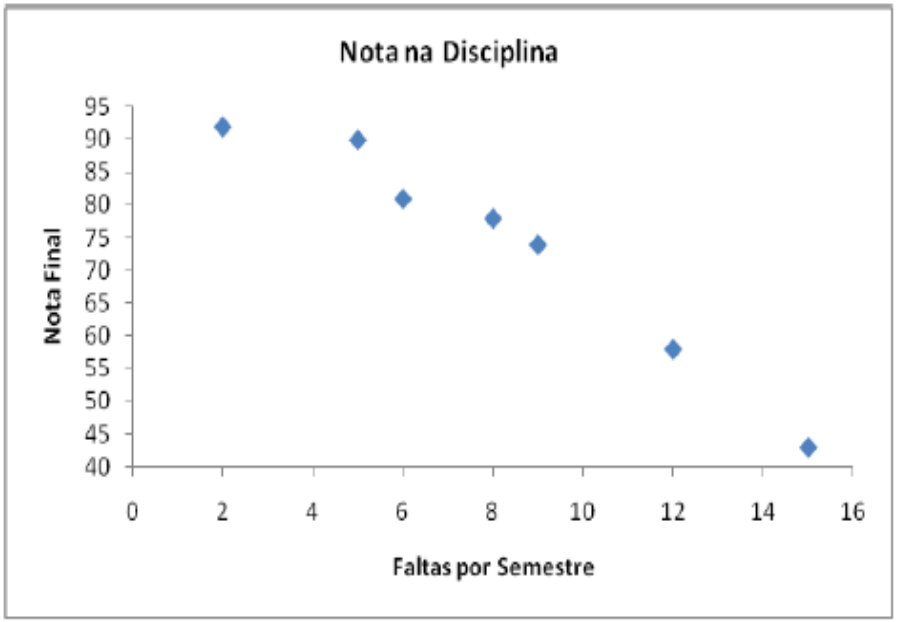
$$r = -0.975$$

***r² = 0.9501633*** (ou 95%)

# 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 |




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

$$r = -0.975$$

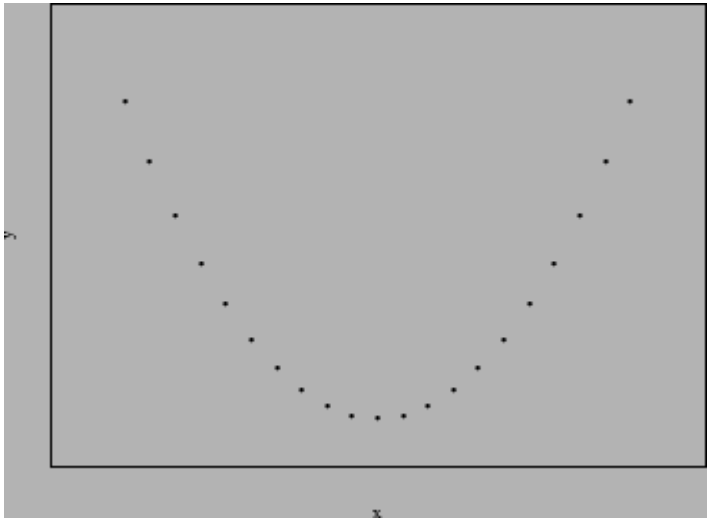
***r² = 0.9501633*** (ou 95%)

*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.*

# Linearidade e normalidade

# Linearidade e normalidade

**Somente relações lineares são detectadas pelo coeficiente de correlação que acabados de descrever** (também chamado de coeficiente de correlação de Pearson)**.**



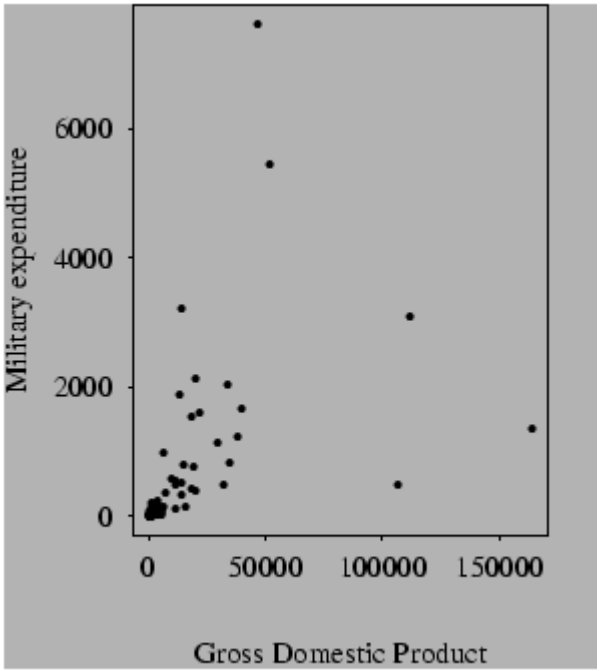
*Aqui o coef. De correlação é igual a zero, mas vemos que existe uma clara relação não-linear entre ambas as variáveis.*



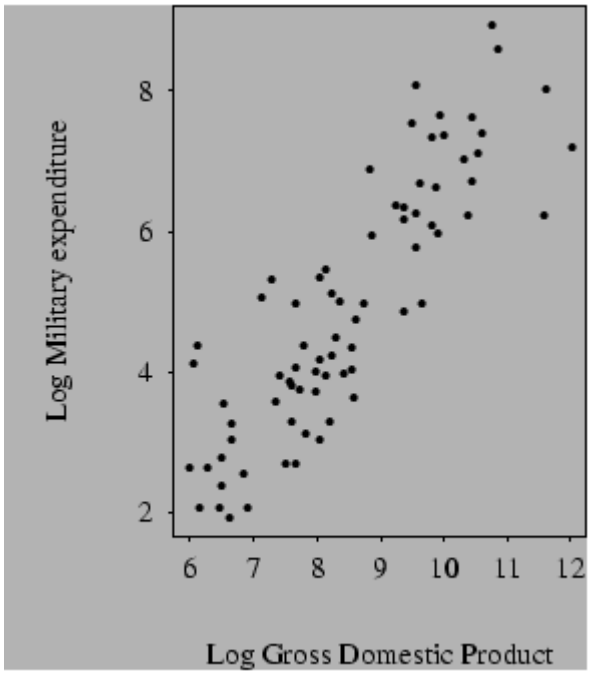
***Dica:*** *Faça gráficos dos dados para visualizar tais relações.*

# Linearidade e normalidade

Em alguns casos pode ser apropriado transformar x e/ou y.

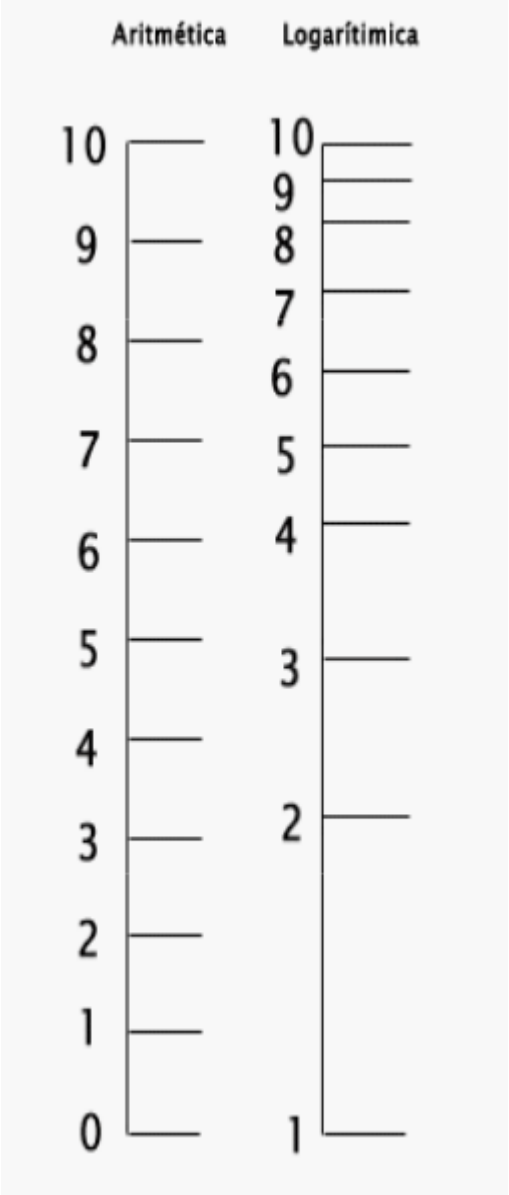



**Escala linear**

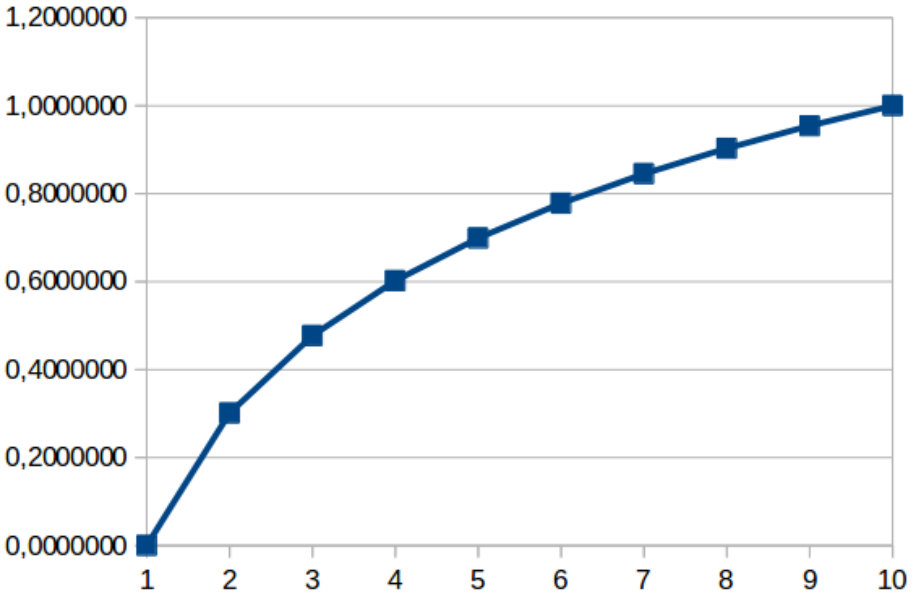



**Escala log-log**

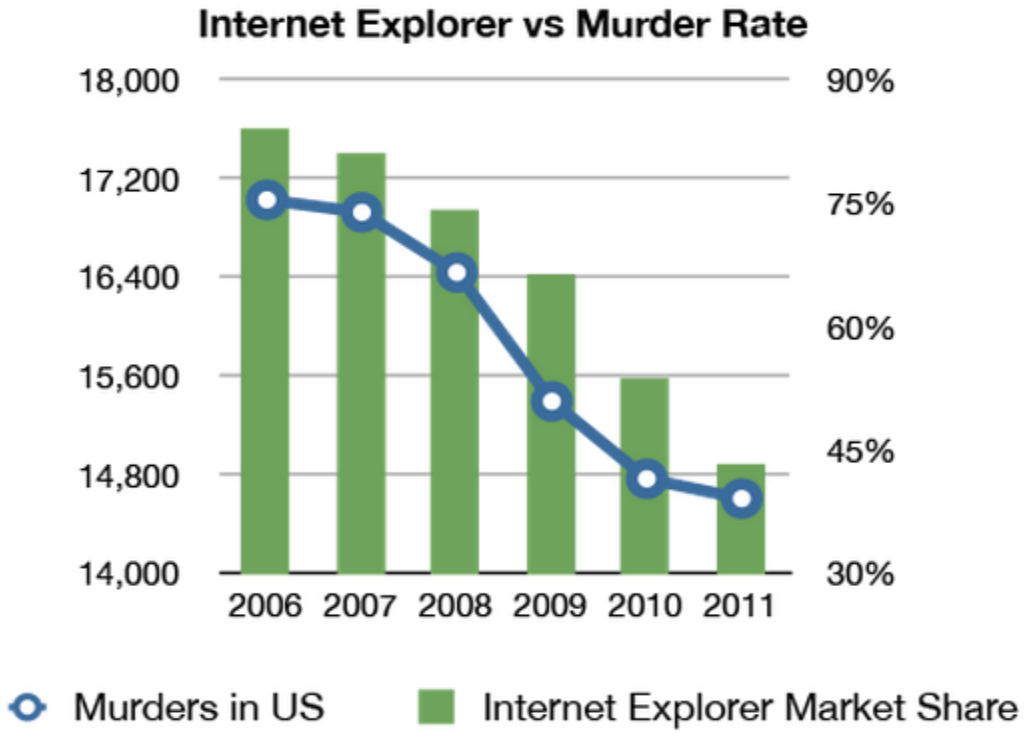
# Linearidade e normalidade




| N | Log(N) |
|---|---|
| 1 | 0,0000000 |
| 2 | 0,3010300 |
| 3 | 0,4771213 |
| 4 | 0,6020600 |
| 5 | 0,6989700 |
| 6 | 0,7781513 |
| 7 | 0,8450980 |
| 8 | 0,9030900 |
| 9 | 0,9542425 |
| 10 | 1,0000000 |



# Causalidade?



# Causalidade e correlação

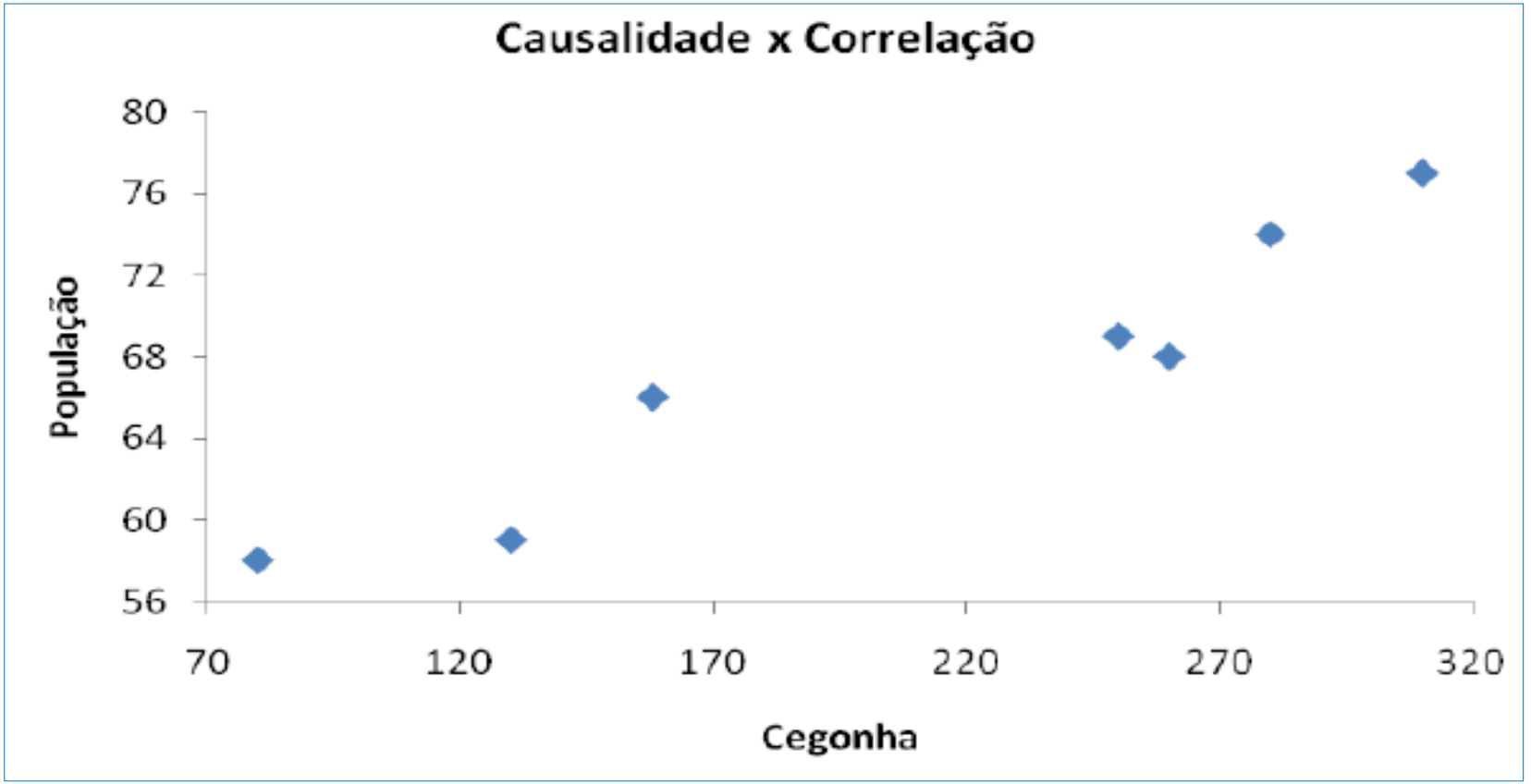
**Correlação não necessariamente implica em causalidade.**

- 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**

# Causalidade e correlação

**Correlação não necessariamente implica em causalidade.**



# Causalidade e correlação

**Correlação não necessariamente implica em causalidade.**

*O número de pessoas usando óculos-de-sol e a quantidade de sorvete consumido em um particular dia são altamente correlacionados.*

*Isto não significa que usar óculos-de-sol causa a compra de sorvetes ou vice-versa!*

**É extremamente difícil estabelecer relações causais a partir de dados observacionais. Precisamos realizar experimentos para obter mais evidências de uma relação causal.**

# 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:

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

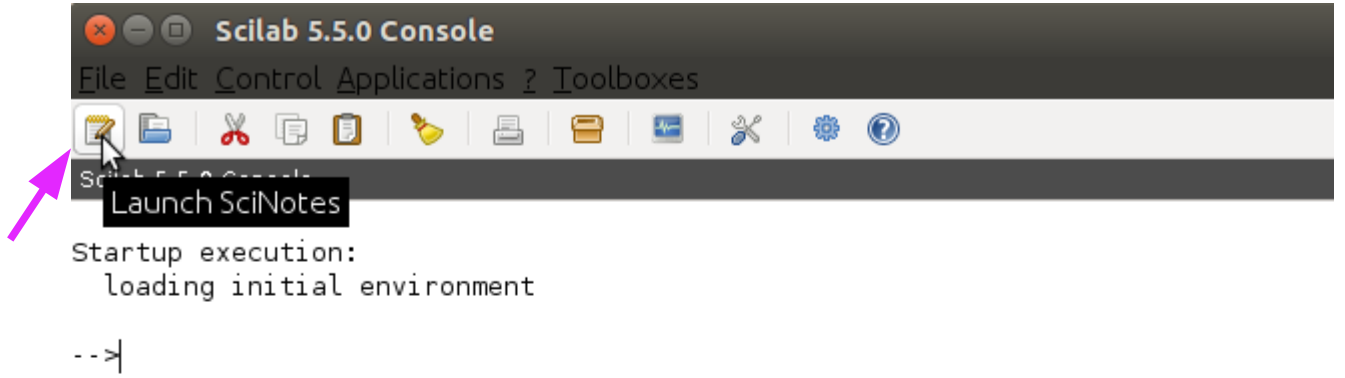
```
-->X = [8, 2, 5, 12, 15, 9, 6];
-->Y = [78, 92, 90, 58, 43, 74, 81];
-->[a,b] = reglin(X,Y)
b =
    1.6
a =
    3.2571429
```

# Criando scripts

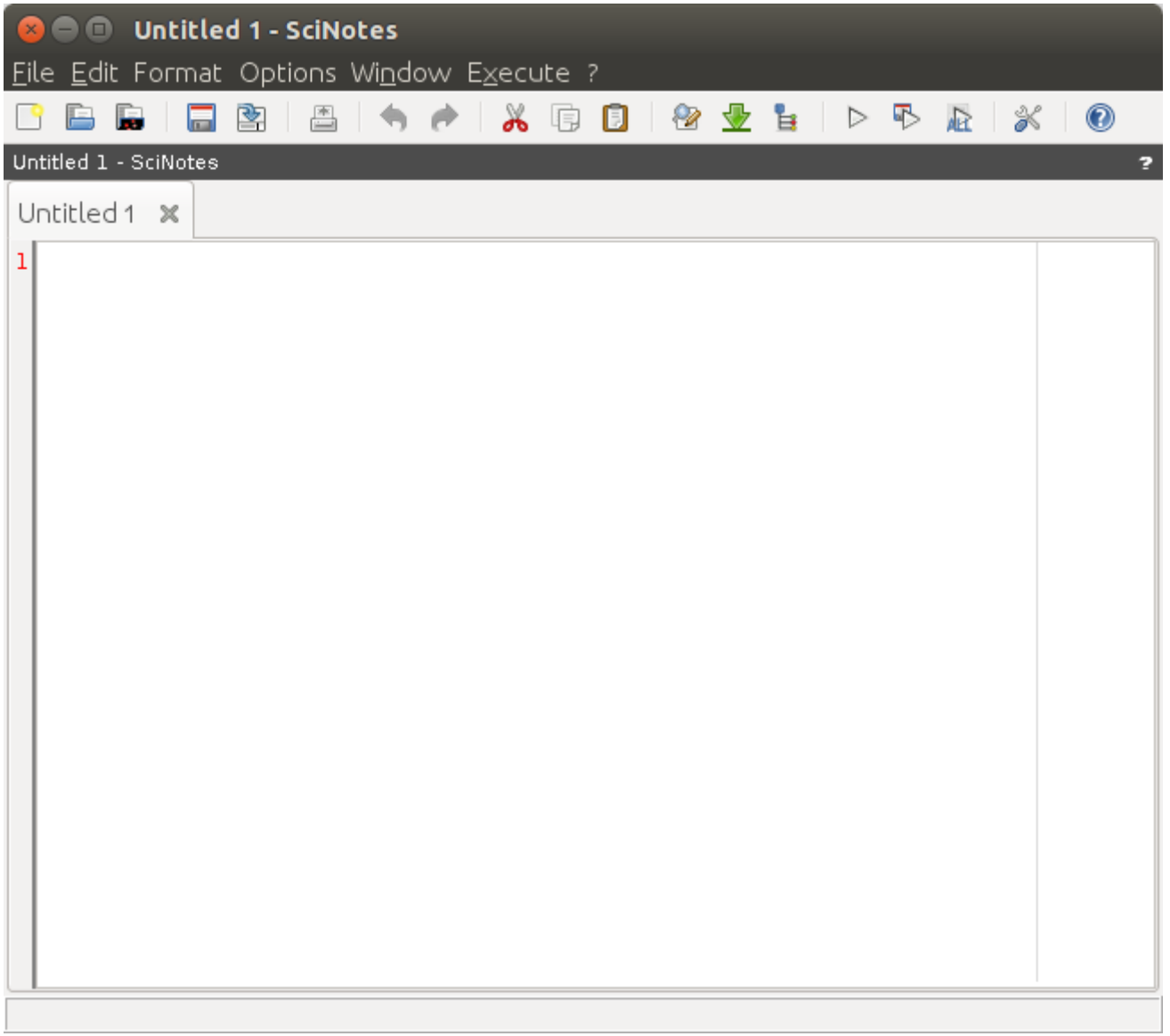
# Criando Scripts

- Até agora estávamos digitando os comandos direto no prompt do scilab
- Uma outra maneira de interagir com o scilab é criar um script, em que uma série de comandos é incluída, pare serem executados em sequência

# Criação de Scripts



# Criação de scripts



# 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 |

Vamos fazer um script que:

- Faça o gráfico destes dados com número total de operações no eixo x.
- Calcule o coeficiente de correlation (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?
- Calcule a reta de regressão, imprima seus coeficientes, e a adicione ao gráfico