Correlação e Regressão Linear com Pandas

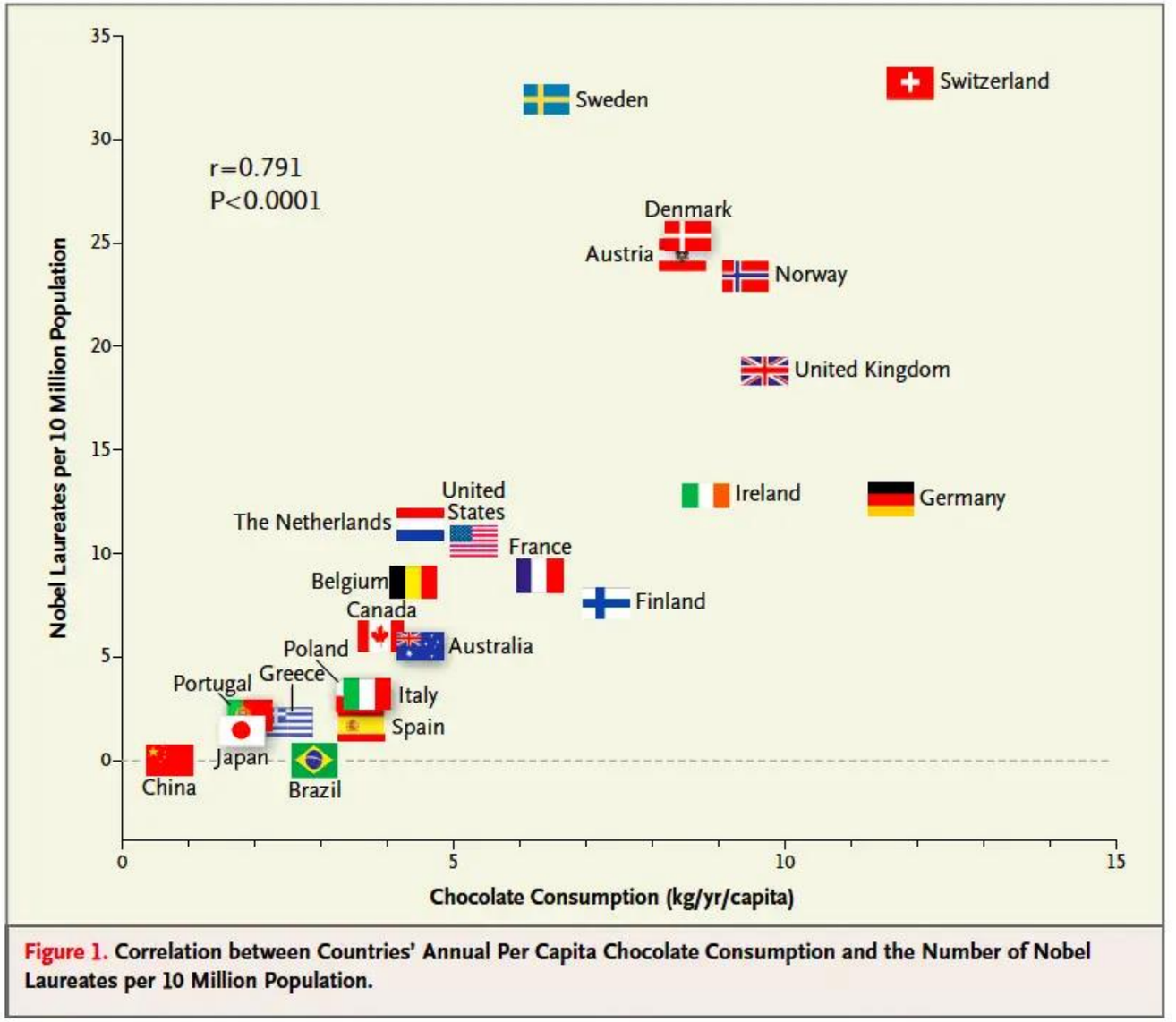


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

https://www.nejm.org/doi/full/10.1056/NEJMon1211064

Professor: Alex Pereira

Coeficiente de Correlação

- Mede a direção e a força de uma relação linear

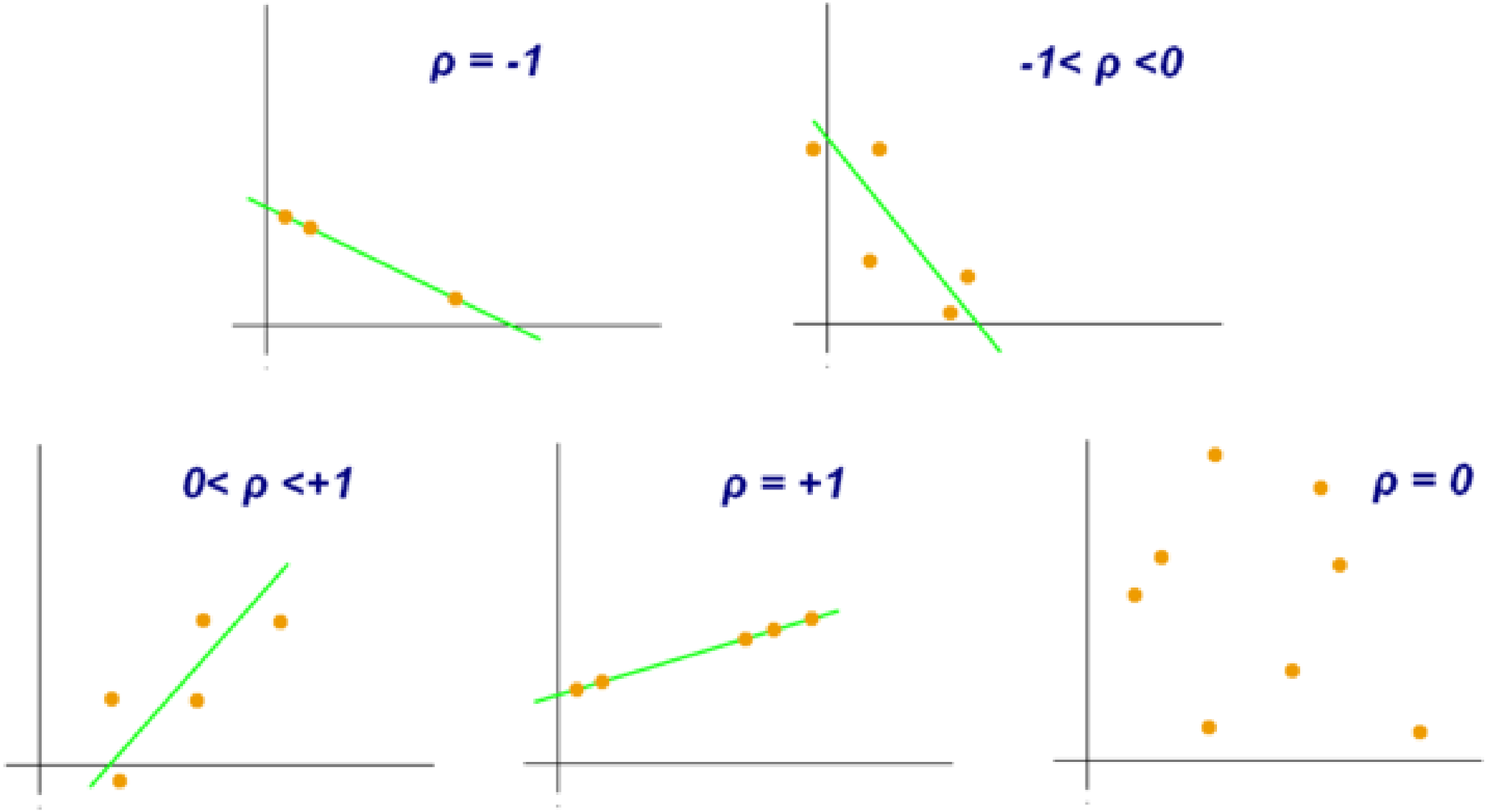


Imagem: https://upload.wikimedia.org/wikipedia/commons/thumb/3/34/Correlation_coefficient.png/600px-Correlation_coefficient.png

Cálculo do Coeficiente de Correlação

A	B	C	D	E	F
Aluno	Altura (cm)	Peso (kg)	Altura em unidade-padrão	Peso em unidade-padrão	(Peso em unidades-padrão) × (Altura em unidades-padrão)*
Nick	185	88	1,34	1,05	1,41
Elana	165	60	−0,49	−0,74	0,36
Dinah	170	70	−0,03	−0,09	0,01
Rebecca	172	67	0,15	−0,29	−0,04
Ben	183	80	1,16	0,54	0,63
Charu	175	58	0,43	−0,87	−0,37
Sahar	150	45	−1,86	−1,69	3,14
Maggie	158	58	−1,13	−0,87	0,98
Faisal	168	77	−0,21	0,35	−0,07
Ted	175	83	0,43	0,73	0,31
Narciso	175	81	0,43	0,61	0,26
Katrina	175	54	0,43	−1,12	−0,48
CJ	187	103	1,52	2,01	3,05
Sophia	155	53	−1,41	−1,18	1,67
Will	185	96	1,34	1,56	2,09
Média	170,34	71,53			Total = 12,95
Desvio padrão	10,91	15,66		Coeficiente de correlação = Total/n = $^{12,95}/_{15} = 0,86$	

Medição em unidade padrão é a distância à média medida em termos de desvio padrão. Calcula-se assim:

$$\frac{x - \bar{x}}{\sigma}$$

Fórmula:

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

Interpretação da Correlação

- Há alta correlação quando
  - a distância à média de uma variável oscilar consistentemente com a distância da outra variável.
- Não depende de unidade de medida
  - Pode-se medir a correlação entre
    - ✓ Peso e altura;
    - ✓ Quantidade de televisores em casa e o desempenho no ensino médio.
  - Isso porque o score padronizado (z-score)
    - ✓ é adimensional

$$z_i = \frac{x_i - \bar{x}}{\sigma}$$

Dataset tips (gorjetas)

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

Cálculo da Correlação com o Pandas

```
1  # Calcule a correlação entre a
2  tips.corr()
```

	total_bill	tip	size
total_bill	1.000000	0.675734	0.598315
tip	0.675734	1.000000	0.489299
size	0.598315	0.489299	1.000000

Mapa de Calor da Correlação

```
3  sns.heatmap(tips.corr(), annot=True)
```

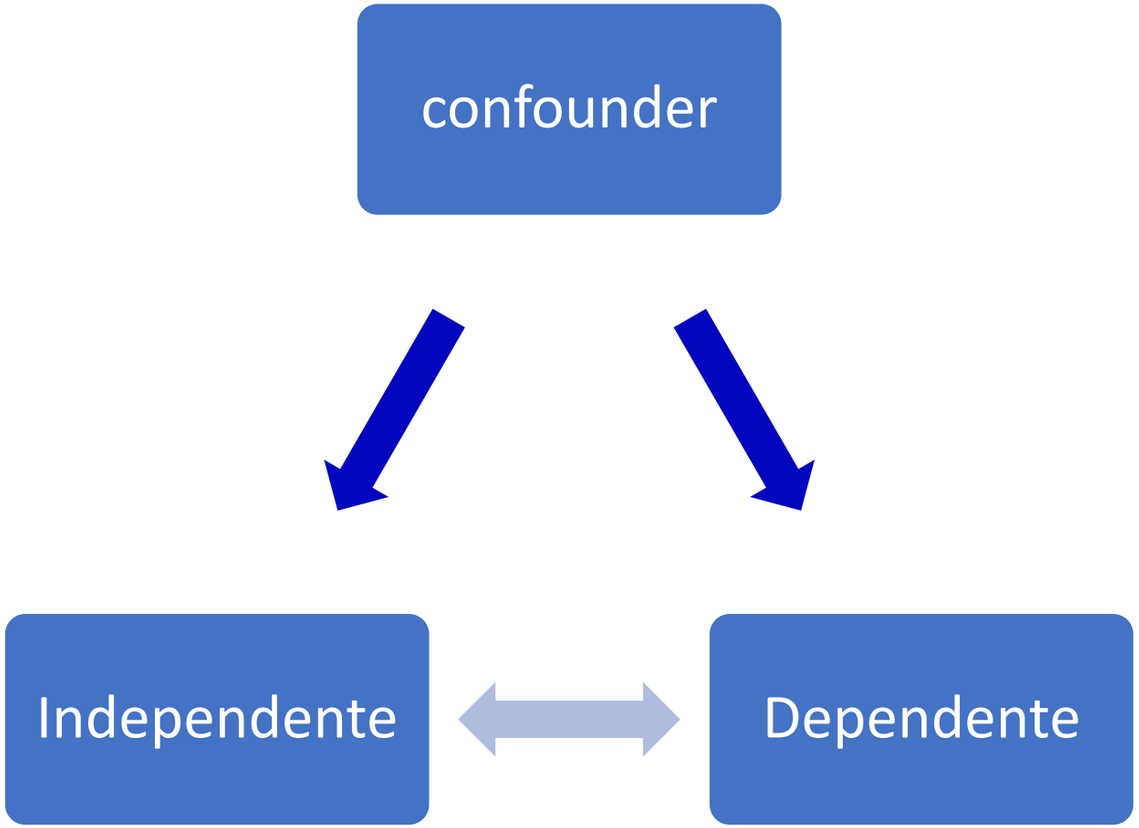
```
<AxesSubplot:>
```

Variável Dependente e Independente

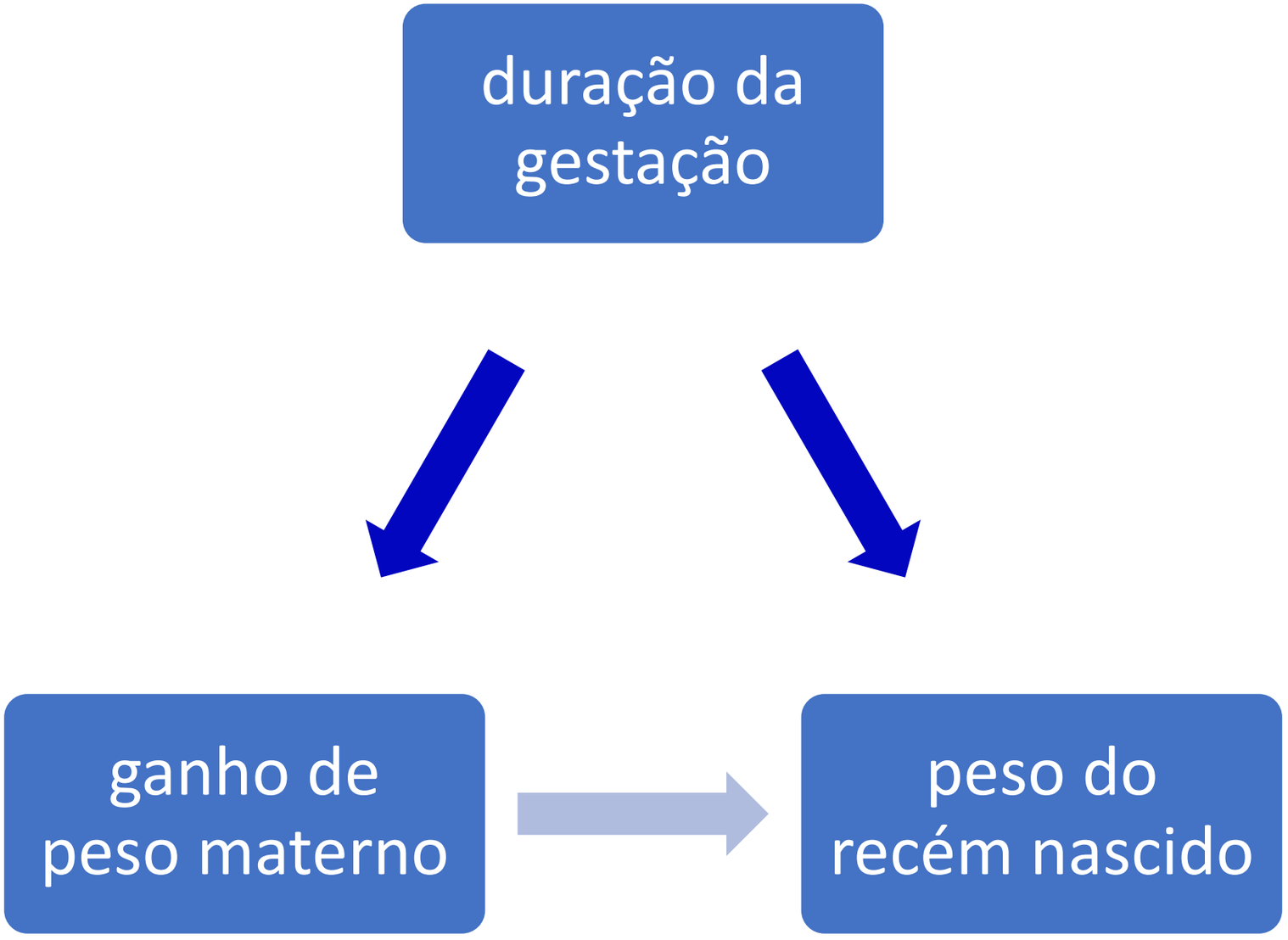
- Variável Independente
  - É a variável que mede uma grandeza que está sendo manipulada num experimento, ou seja,
    - ✓ é aquela que você está estudando os seus efeitos sobre outras variáveis.
  - Sinônimos: variáveis explicativas ou de controle.
- Variável Dependente
  - É a que está sendo explicada,
    - ✓ que depende de outros fatores
- Exemplos:
  - Viscosidade e o atrito de uma esfera em queda livre num líquido;
  - Consumo percapita de chocolate e prêmios nobel.

Confounding Variable (Variável de confusão)

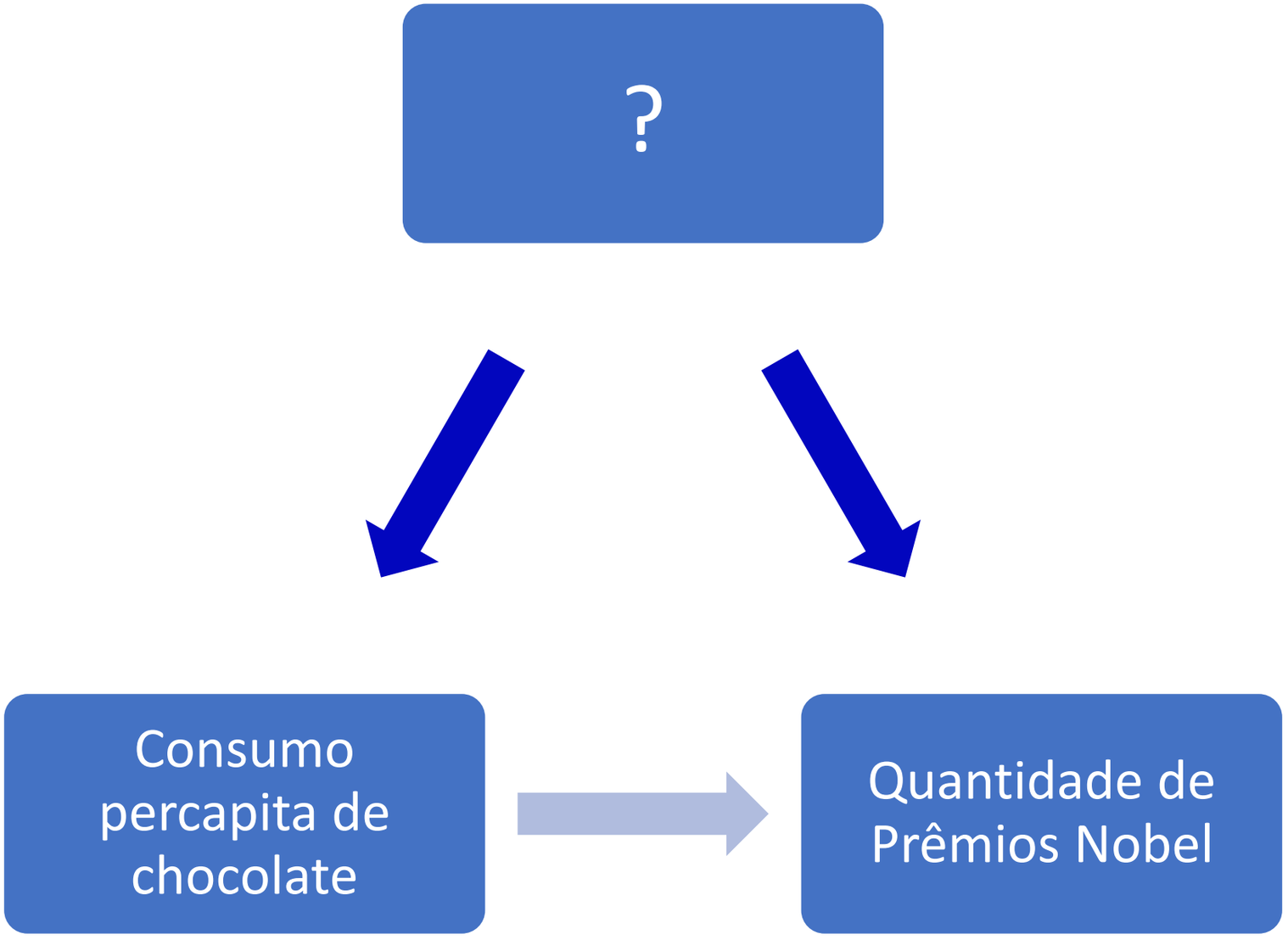
- Em investigação de causa e efeito,
  - uma variável de confusão é uma terceira variável não medida
    - ✓ que influencia tanto a suposta causa quanto o suposto efeito.



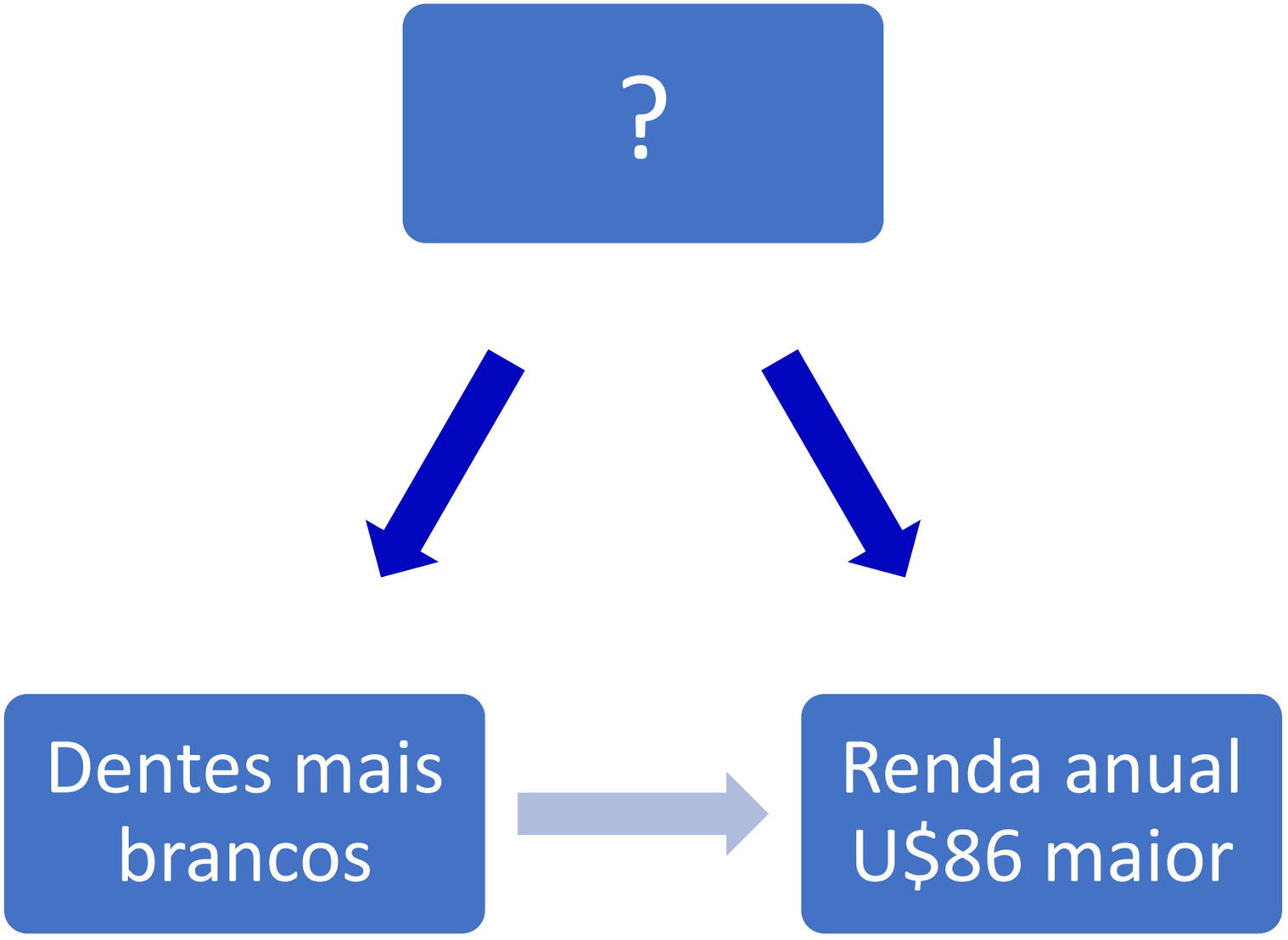
Confounding Variable - Exemplo



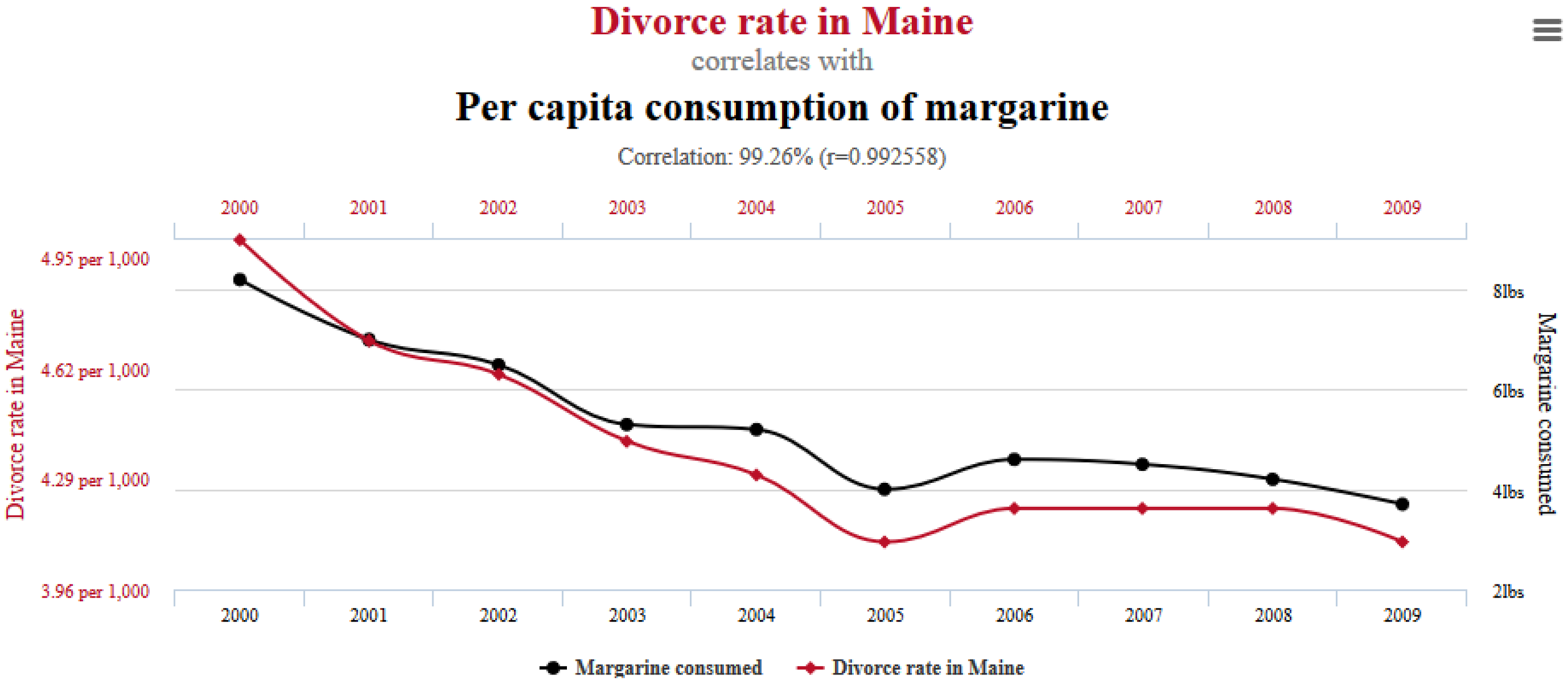
Confounding Variable – Exemplo (2)



Confounding Variable – Exemplo (2)

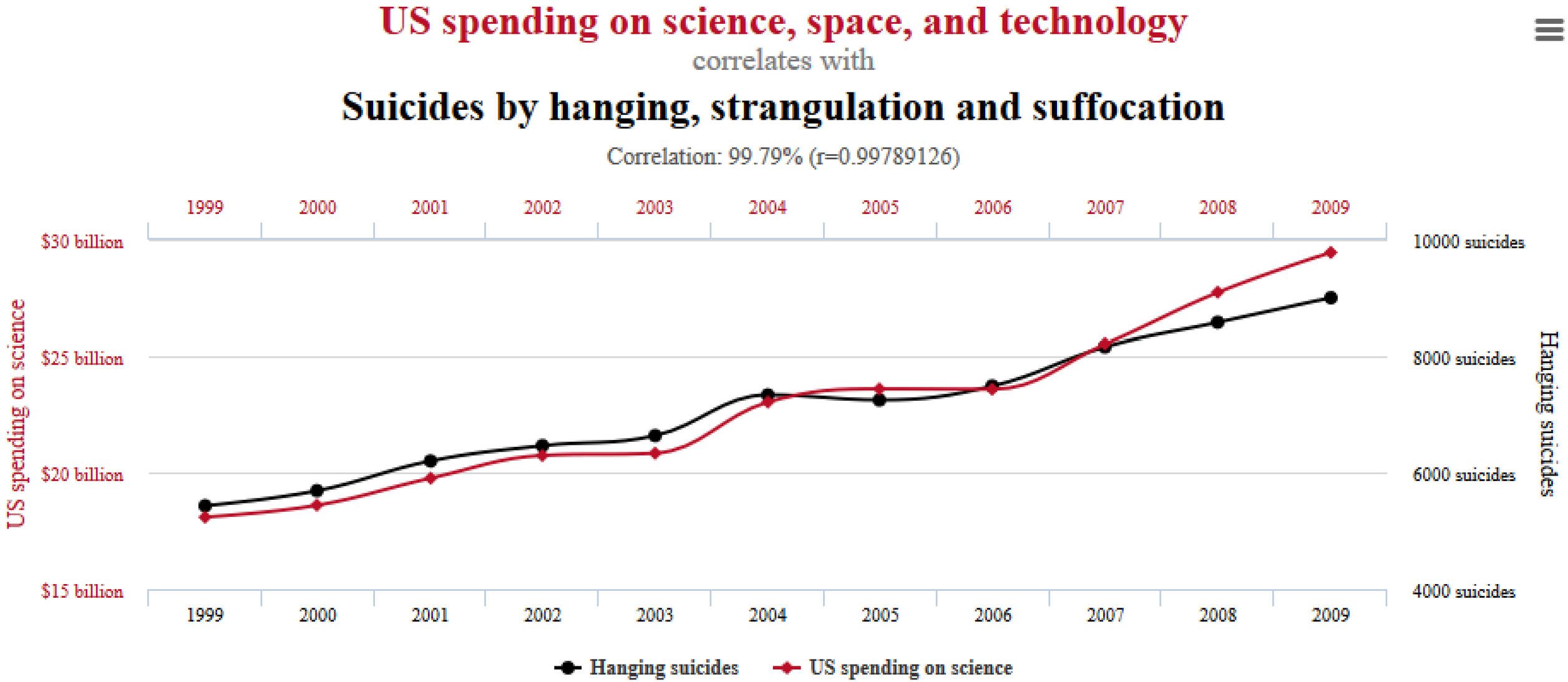


Outros exemplos de correlações espúrias



Fonte: https://www.tylervigen.com/spurious-correlations

Outros exemplos de correlações espúrias (2)



Fonte: https://www.tylervigen.com/spurious-correlations

Regressão Linear

Baixo Controle e Doenças Cardíacas

- Num estudo longitudinal com servidores públicos britânicos
  - Trabalhadores com pouco controle sobre suas responsabilidades
    - ✓ o que significa que têm pouco a dizer sobre quais serviços executar ou como são executados
  - Têm uma taxa de mortalidade maior do que outros servidores com maior autoridade na tomada de decisões.
- Não é o estresse associado à responsabilidades importantes
  - que é mais prejudicial,
    - ✓ é o estresse associado a lhe dizerem o que fazer enquanto você tem pouco a dizer sobre como e quando fazer.
- Como se chegou a essa conclusão?
  - Eliminando *confounding variables* e usando análise de regressão
    - ✓ Educação, hábitos de fumar

Metodologia de Cálculo da Regressão Linear

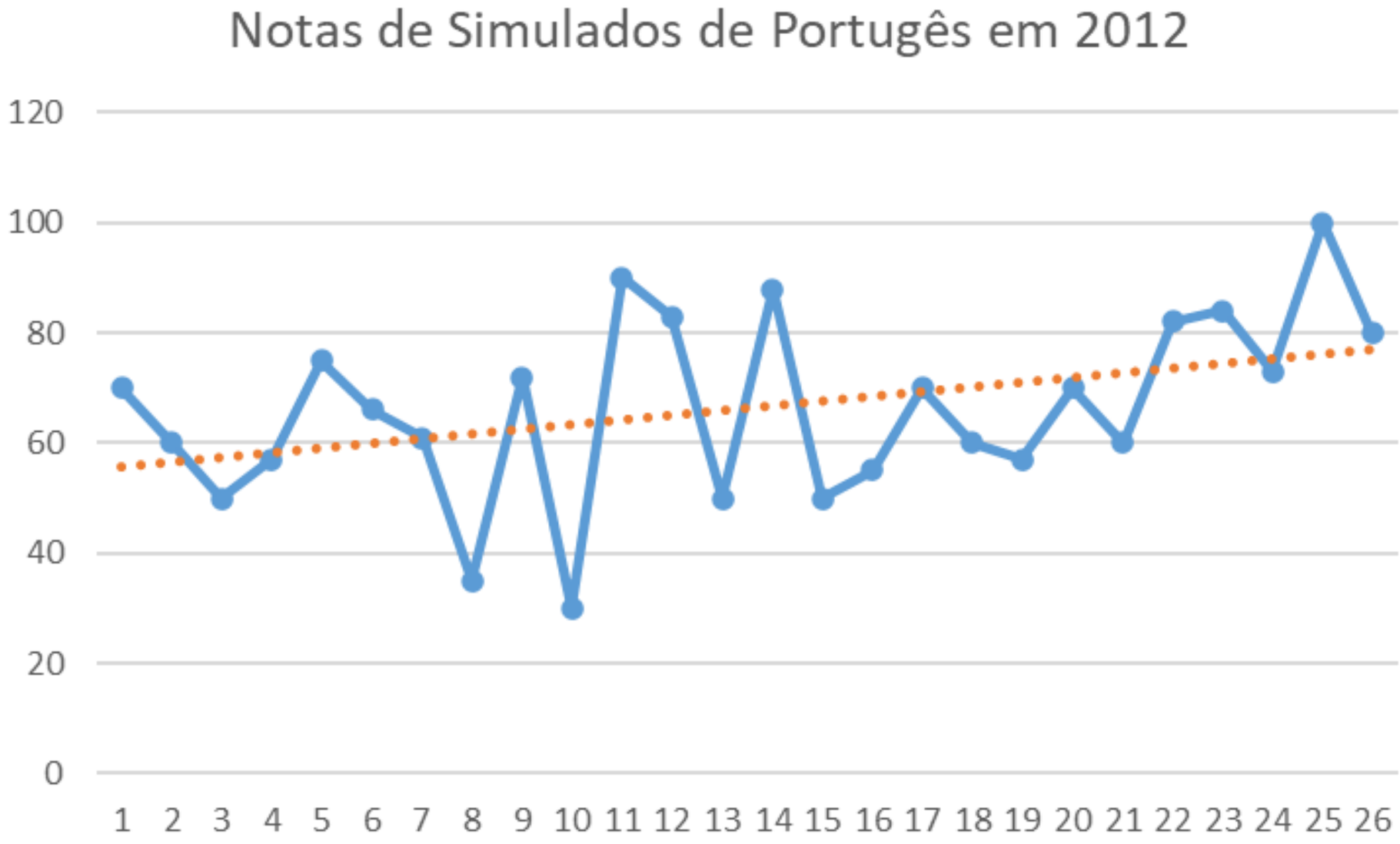
Fonte: WHEELAN, C. Estatística, o que é, para que serve, como funciona.

Quais valores de a e b que minimizam a soma dos resíduos ?

$$y = ax + b$$

Um método chamado Mínimos Quadrados Ordinários encontra a resposta.

Regressão como ferramenta de Administração/Controle



Coeficiente de Regressão (Coef)

- Tamanho do efeito observado representado pelo
  - é o parâmetro *A* na equação

$$y = Ax + B$$

- Exemplo:

$$PESO = A \times ALTURA + B$$

$$PESO = 0{,}8 \times ALTURA - 60$$

3537 amostras (Changing Lives)

* ALTURA em centímetros

R^2 da Regressão Linear

- É uma medida do tamanho total da variação
  - explicado pela equação de regressão
    - ✓ varia de 0 a 1
- Quanto maior o valor de R^2
  - mais explicativo é o modelo
    - ✓ ou a variável independente (no caso de uma regressão simples)

Erro padrão do Coeficiente de Regressão Linear

- É uma medida da dispersão dos valores do coeficiente de regressão (Coef)
  - ou seja, o desvio padrão desses valores
    - ✓ representa a distância média que os valores observados desviam da linha de regressão.
  - S ou SE (Standard Error)
- Diferentemente do R^2
  - o erro padrão (SE) pode ser usado para avaliar a precisão das predições
  - Aproximadamente 95% das observações devem cair dentro do intervalo
    - ✓ ± 2 * SE

$$y = Ax + B$$

Outras estatísticas da Regressão Linear

- P-Value
  - a probabilidade de se obter
- Estatística t
  - Calculado por: SE / p-value
- Intervalo de confiança medido em termos do erro padrão
  - Intervalo de 95%

$$Coef \pm 2 * SE$$

Regressão linear com Statsmodel (Python)

- Boston House Prices dataset

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
LSTAT	% lower status of the population
MEDV	**Median value of owner-occupied homes in $1000's**

Tratamento de Exceção em Python

- Divisão por zero gera uma exceção
  - Como fazer para o seu programa continuar a execução,
    - ✓ e falhar graciosamente ?

```
2/2
```

```
1.0
```

```
a = 0 / 0
print("Passou aqui")
```

```
---------------------------------------------------------------------------
ZeroDivisionError                         Traceback (most recent call last)
<ipython-input-23-efda217e1f08> in <module>
----> 1 a = 0 / 0
      2 print("Passou aqui")

ZeroDivisionError: division by zero
```

Tratamento de Exceção em Python

- Use a sintaxe de tratamento de exceção com
  - try e except, conforme o exemplo a seguir
- Boas práticas
  - Coloque poucas linhas de código dentro do escopo do try/except
  - Capture exceções específicas

```
try:
    a = 0/0
    print("resultado: {0}".format(i))
except ZeroDivisionError as e:
    print(e)
    print("deu erro.")
    pass
```

```
division by zero
deu erro.
```

Prática no Jupyter Notebook

- Faça os exercícios da aula;
- Há exercícios extra.