

# Testes de hipóteses

Wagner H. Bonat  
Fernando P. Mayer  
Elias T. Krainski

Universidade Federal do Paraná  
Departamento de Estatística  
Laboratório de Estatística e Geoinformação

13/06/2018



# Sumário

1 Introdução

2 Regressão linear

3 Correlação

# Introdução

Um problema comum é o estudo da relação entre duas variáveis,  $X$  e  $Y$ . Na prática, procura-se uma **função** de  $X$  que explique  $Y$ , ou seja,

$$X, Y \rightarrow Y \simeq f(X)$$

Essa relação, em geral, não é perfeita, ou seja, existem **erros** associados.

# Introdução

Uma das preocupações estatísticas ao analisar dados é a de criar **modelos** do fenômeno em observação.

As observações frequentemente estão misturadas com variações **acidentais** ou **aleatórias**.

Assim, é conveniente supor que cada observação é formada por duas partes: uma **previsível** (ou controlada) e outra **aleatória** (ou não previsível), ou seja

$$(\text{observação}) = (\text{previsível}) + (\text{aleatório})$$

# Introdução

$$(\text{observação}) = (\text{previsível}) + (\text{aleatório})$$

A parte previsível, incorpora o conhecimento sobre o fenômeno, e é usualmente expressa por uma **função matemática** com **parâmetros desconhecidos**.

A parte aleatória deve obedecer algum **modelo de probabilidade**

Com isso, o trabalho é produzir **estimativas** para os parâmetros desconhecidos, com base em amostras observadas.

# Introdução

$$(\text{observação}) = (\text{previsível}) + (\text{aleatório})$$

Simbolicamente, podemos escrever

$$y_i = \theta + e_i$$

onde -  $y_i$  = observação  $i$  -  $\theta$  = efeito fixo, comum a todos os indivíduos -  $e_i$  = “erro” da observação  $i$ , ou efeito residual ou aleatório

$e_i$  pode ser considerado como o efeito resultante de várias características que não estão explícitas no modelo.

# Introdução

**Exemplo:** considerando que o peso médio da população é de  $\mu 62$  kg, então o peso de cada pessoa  $y_i$  pode ser descrita pelo seguinte modelo

$$y_i = 62 + e_i$$

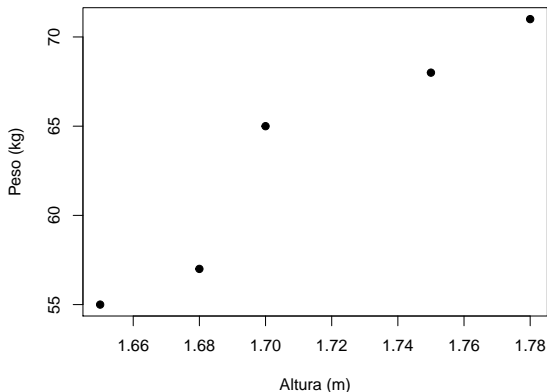
onde  $\theta = \mu$ , e cada  $e_i$  determinará o peso de cada pessoa, em função de diversos fatores como: altura, sexo, idade, país, . . . , ou seja

$$e_i = f(\text{altura, sexo, idade, país, } \dots)$$

Ou seja, à medida que **relacionamos** o peso com outras variáveis, ganhamos informação e diminuimos o **erro**.

# Introdução

Por exemplo, podemos relacionar os pesos de 5 pessoas com suas respectivas alturas.

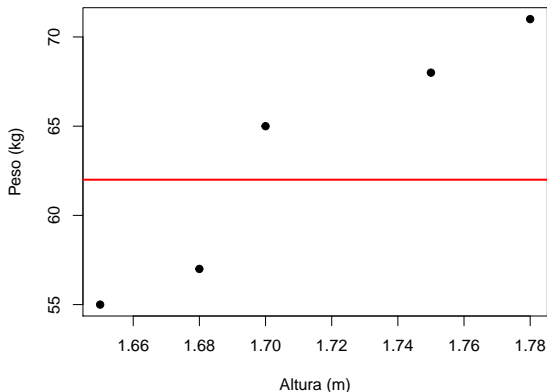


E notamos que existe uma aparente **relação linear** entre estas duas variáveis



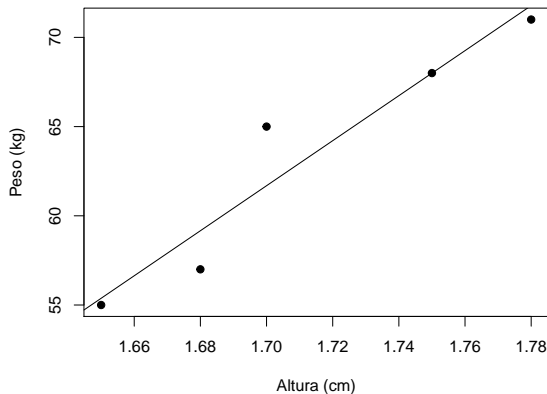
# Introdução

Por exemplo, podemos relacionar os pesos de 5 pessoas com suas respectivas alturas.



E notamos que existe uma aparente **relação linear** entre estas duas variáveis.

# Introdução



Como o peso depende da altura de maneira linear, podemos então aprimorar o modelo anterior incorporando essa informação.

# Introdução

Um **modelo linear** entre duas variáveis  $X$  e  $Y$ , é definido matematicamente como uma equação com dois parâmetros desconhecidos,

$$Y = \beta_0 + \beta_1 X$$

Sendo assim, o modelo anterior onde conhecíamos só a média  $\mu$ ,

$$y_i = \mu + e_i$$

pode ser reescrito como

$$y_i = \beta_0 + \beta_1 \text{ altura} + e_i$$

Note que o erro deve diminuir, pois agora

$$e_i = f(\text{altura, sexo, idade, país, } \dots)$$

ou seja, incorporamos uma informação para explicar o peso, que antes estava inserida no erro.

# Introdução

No exemplo anterior, notamos que o peso é uma variável **dependente** (linearmente) da altura.

A **análise de regressão** é a técnica estatística que analisa as relações existentes entre uma única variável **dependente**, e uma ou mais variáveis **independentes**.

O objetivo é estudar as relações entre as variáveis, a partir de um **modelo matemático**, permitindo **estimar** o valor de uma variável a partir da outra.

- Exemplo: sabendo a altura podemos determinar o peso de uma pessoa, se conhecemos os parâmetros do modelo anterior

# Introdução

O problema da análise de regressão consiste em definir a **forma** de relação existente entre as variáveis.

Por exemplo, podemos ter as seguintes relações

$$Y = \beta_0 + \beta_1 X \quad \text{linear}$$

$$Y = \beta_0 X^{\beta_1} \quad \text{potência}$$

$$Y = \beta_0 e^{\beta_1 X} \quad \text{exponencial}$$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 \quad \text{polinomial}$$

Em todos os casos, a variável **dependente** é  $Y$ , aquela que será **predita** a partir da relação e da variável **independente**  $X$

# Sumário

1 Introdução

2 Regressão linear

3 Correlação

# Regressão linear

Em uma **análise de regressão linear** consideraremos apenas as variáveis que possuem uma **relação linear** entre si.

Uma análise de regressão linear **múltipla** pode associar  $k$  variáveis independentes ( $X$ ) para “explicar” uma única variável dependente ( $Y$ ),

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e$$

Uma análise de regressão linear **simples** associa uma única variável independente ( $X$ ) com uma variável dependente ( $Y$ ),

$$Y = \beta_0 + \beta_1 X + e$$

# Regressão linear

Assim, dados  $n$  pares de valores,  $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ , se for admitido que  $Y$  é função linear de  $X$ , pode-se estabelecer uma regressão linear simples, cujo modelo estatístico é

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad i = 1, 2, \dots, n$$

onde: -  $Y$  é a variável **resposta** (ou **dependente**) -  $X$  é a variável **explicativa** (ou **independente**) -  $\beta_0$  é o **intercepto** da reta (valor de  $Y$  quando  $X = 0$ ) -  $\beta_1$  é o **coeficiente angular** da reta (efeito de  $X$  sobre  $Y$ ) -  $e \sim N(0, \sigma^2)$  é o **erro**, ou **desvio**, ou **resíduo**

O problema agora consiste em **estimar** os parâmetros  $\beta_0$  e  $\beta_1$ .



# Interpretação dos parâmetros

$\beta_0$  representa o ponto onde a reta corta o eixo  $Y$  (na maioria das vezes não possui interpretação prática)

$\beta_1$  representa a variabilidade em  $Y$  causada pelo aumento de uma unidade em  $X$ . Além disso,

- $\beta_1 > 0$  mostra que com o aumento de  $X$ , também há um aumento em  $Y$
- $\beta_1 = 0$  mostra que **não há efeito** de  $X$  sobre  $Y$
- $\beta_1 < 0$  mostra que com a aumento de  $X$ , há uma diminuição em  $Y$

# Estimação dos parâmetros

Como através de uma amostra obtemos uma estimativa da verdadeira equação de regressão, denominamos

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

ou seja,  $\hat{Y}_i$  é o valor **estimado** de  $Y_i$ , através das **estimativas** de  $\beta_0$  e  $\beta_1$ , que chamaremos de  $\hat{\beta}_0$  e  $\hat{\beta}_1$ .

Para cada valor de  $Y_i$ , temos um valor  $\hat{Y}_i$  estimado pela equação de regressão,

$$Y_i = \hat{Y}_i + e_i$$

# Estimação dos parâmetros

Portanto, o erro (ou desvio) de cada observação em relação ao modelo adotado será

$$e_i = Y_i - \hat{Y}_i$$

$$e_i = Y_i - (\beta_0 + \beta_1 X_i)$$

Devemos então adotar um modelo cujos parâmetros  $\beta_0$  e  $\beta_1$ , tornem esse diferença a menor possível.

Isso equivale a **minimizar a soma de quadrados dos resíduos (SQR)**, ou do erro,

$$SQR = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2$$

## Estimação dos parâmetros

O método de minimizar a soma de quadrados dos resíduos é denominado de **método dos mínimos quadrados**.

Para se encontrar o ponto mínimo de uma função, temos que obter as derivadas parciais em relação a cada parâmetro,

$$\frac{\partial SQR}{\partial \beta_0} = 2 \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_i](-1)$$
$$\frac{\partial SQR}{\partial \beta_1} = 2 \sum_{i=1}^n [Y_i - \beta_0 - \beta_1 X_i](-X_i)$$

e igualar os resultados a zero

$$\hat{\beta}_0 = \frac{\partial SQR}{\partial \beta_0} = 0 \quad \text{e} \quad \hat{\beta}_1 = \frac{\partial SQR}{\partial \beta_1} = 0$$

# Estimação dos parâmetros

Dessa forma, chegamos às **estimativas de mínimos quadrados** para os parâmetros  $\beta_0$  e  $\beta_1$ :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

onde

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{e} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

# Exemplo

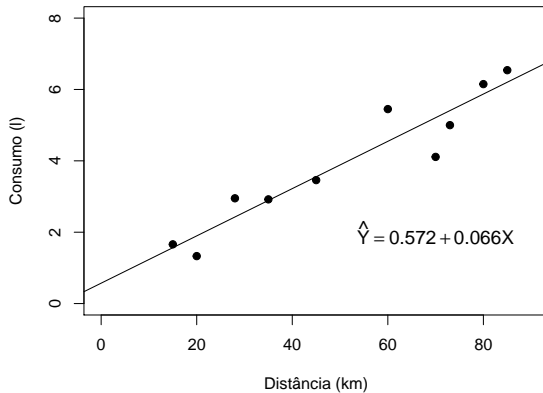
A tabela a seguir relaciona as distâncias percorridas por carros (km) e seus consumos de combustível (litros), em uma amostra de 10 carros novos.

| Distância (km) | Consumo (litros) |
|----------------|------------------|
| 20             | 1,33             |
| 60             | 5,45             |
| 15             | 1,66             |
| 45             | 3,46             |
| 35             | 2,92             |
| 80             | 6,15             |
| 70             | 4,11             |
| 73             | 5,00             |
| 28             | 2,95             |
| 85             | 6,54             |

Com isso:

- a. Faça um diagrama de dispersão
- b. Traçe um modelo linear aproximado
- c. Estime os parâmetros  $\hat{\beta}_0$  e  $\hat{\beta}_1$
- d. Interprete o resultado. Pode-se concluir que para percursos mais longos há maior consumo de combustível?
- e. Faça uma *predição* do consumo de combustível para uma distância de 50 km.

## Exemplo



# Exemplo

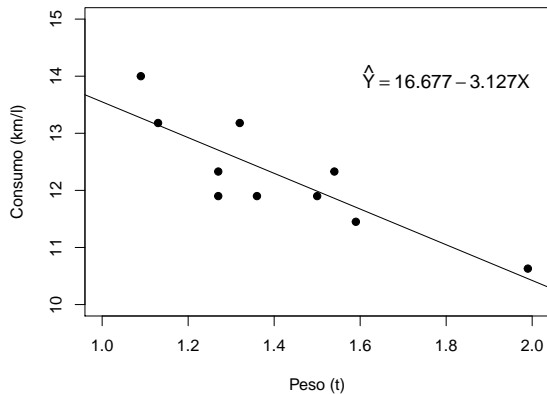
A tabela a seguir relaciona os pesos de carros (t) e as taxas de consumo de combustível (km/l), para uma amostra de 10 carros. |||||Peso & 1,32 & 1,59 & 1,27 & 1,99 & 1,13 & 1,54 & 1,36 & 1,5 & 1,27 & 1,09 \ Consumo & 13,18 & 11,45 & 12,33 & 10,63 & 13,18 & 12,33 & 11,90 & 11,90 & 11,90 & 14,00 \

Com isso:

- Faça um diagrama de dispersão
- Traça um modelo linear aproximado
- Estime os parâmetros  $\hat{\beta}_0$  e  $\hat{\beta}_1$
- Interprete o resultado. O que você pode concluir a respeito da taxa de consumo?
- Faça uma *predição* do consumo de combustível para um veículo com peso de 1,8 t.



## Exemplo



# Sumário

1 Introdução

2 Regressão linear

3 Correlação

# Correlação

Até agora o interesse estava em estudar qual a influência de uma V.A.  $X$  sobre uma V.A.  $Y$ , por meio de uma **relação linear**.

Assim, em uma análise de regressão é indispensável identificar qual variável é dependente.

Na **análise de correlação** isto não é necessário, pois queremos estudar o **grau de relacionamento** entre as variáveis  $X$  e  $Y$ , ou seja, uma medida de **covariabilidade** entre elas.

A correlação é considerada como uma medida de **influência mútua** entre variáveis, por isso não é necessário especificar quem influencia e quem é influenciado.

# Correlação

O grau de relação entre duas variáveis pode ser medido através do coeficiente de correlação linear ( $r$ ), dado por

$$r = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sqrt{\sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}} \cdot \sqrt{\sum_{i=1}^n Y_i^2 - \frac{(\sum_{i=1}^n Y_i)^2}{n}}} r = \frac{\text{Cov}(XY)}{\text{DP}(X) \cdot \text{DP}(Y)}$$

onde

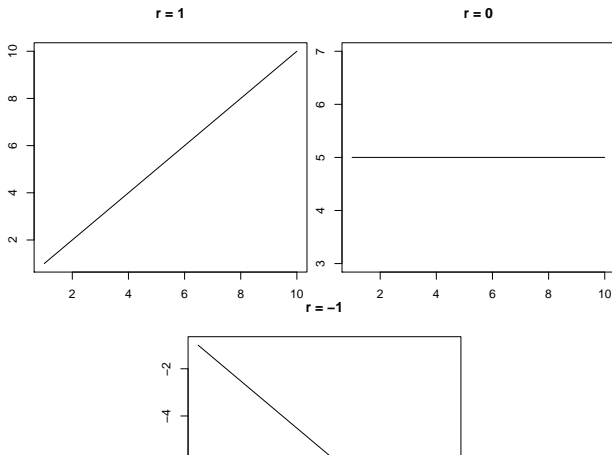
$$-1 \leq r \leq 1$$

Portanto,

- $r = 1$  correlação **positiva** perfeita entre as variáveis
- $r = 0$  **não há** correlação entre as variáveis
- $r = -1$  correlação **negativa** perfeita entre as variáveis

# Correlação

Existem muitos tipos de associações possíveis, e o coeficiente de correlação avalia o quanto uma nuvem de pontos no gráfico de dispersão se aproxima de uma reta.



# Coeficiente de determinação

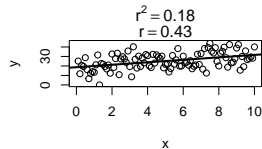
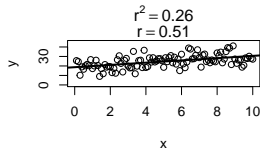
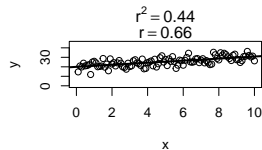
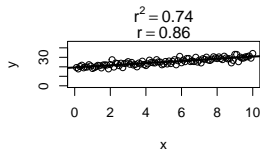
O **coeficiente de determinação** ( $r^2$ ) é o quadrado do coeficiente de correlação, por consequência

$$0 \leq r^2 \leq 1$$

O  $r^2$  nos dá a **porcentagem de variação em  $Y$  que pode ser explicada pela variável independente  $X$ .**

Quanto mais próximo de 1, maior é a explicação da variável  $Y$  pela variável  $X$ .

# Correlação



# Exemplo

Usando os dados do primeiro exemplo anterior, calcule o coeficiente de correlação e o  $r^2$ .

||||||||||||||||

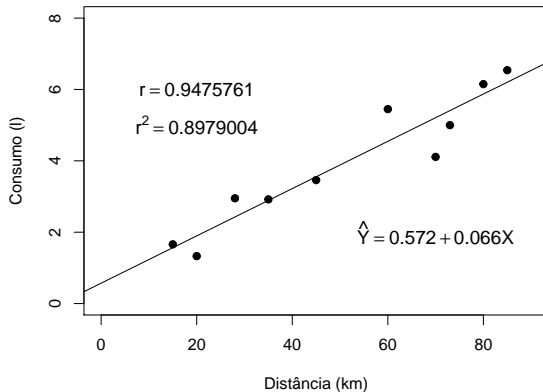
Distância & 20 & 60 & 15 & 45 & 35 & 80 & 70 & 73 & 28 & 85 \ Consumo & 1,33 & 5,45 & 1,66 & 3,46 & 2,92 & 6,15 & 4,11 & 5,00 & 2,95 & 6,54 \

$$\sum_{i=1}^n X_i = 511 \quad \sum_{i=1}^n Y_i = 39.57 \quad \sum_{i=1}^n X_i Y_i = 2419.6 \quad \sum_{i=1}^n X_i^2 = 3.2113 \times 10^4$$

$$\sum_{i=1}^n Y_i^2 = 185.9137$$



## Exemplo



## Exemplo

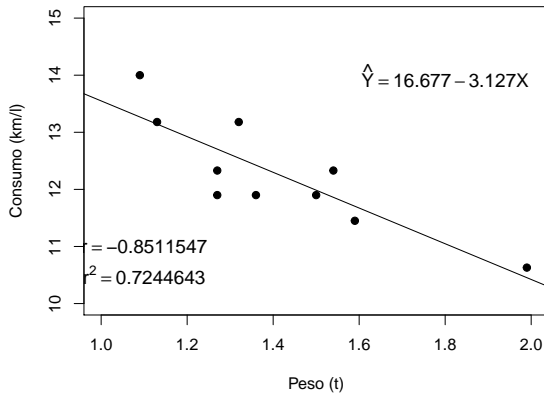
Usando os dados do segundo exemplo anterior, calcule o coeficiente de correlação e o  $r^2$ .

||||||||||||

Peso & 1,32 & 1,59 & 1,27 & 1,99 & 1,13 & 1,54 & 1,36 & 1,5 & 1,27 & 1,09 \ Consumo & 13,18 & 11,45 & 12,33 & 10,63 & 13,18 & 12,33 & 11,90 & 11,90 & 11,90 & 14,00 \

$$\sum_{i=1}^n X_i = 14.06 \quad \sum_{i=1}^n Y_i = 122.8 \quad \sum_{i=1}^n X_i Y_i = 170.7045 \quad \sum_{i=1}^n X_i^2 = 20.3926 \quad \sum_{i=1}^n Y_i^2 = 1516.412$$

## Exemplos



# Teste para o coeficiente de correlação

Usualmente definimos o coeficiente de correlação para uma amostra, pois desconhecemos esse valor para a população.

Uma população que tenha duas variáveis não correlacionadas pode produzir uma amostra com coeficiente de correlação diferente de zero.

Para **testar** se uma amostra foi colhida de uma população para o qual o coeficiente de correlação entre duas variáveis é nulo, precisamos obter a **distribuição amostral** da estatística  $r$ .

# Teste para o coeficiente de correlação

Seja  $\rho$  o verdadeiro coeficiente de correlação populacional desconhecido.

Para testar se o coeficiente de correlação populacional é igual a zero, realizamos um teste de hipótese com

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

A estatística de teste utilizada é

$$t_{calc} = r \sqrt{\frac{n-2}{1-r^2}}$$

que tem distribuição  $t$  de Student com  $n - 2$  graus de liberdade.

# Teste para o coeficiente de correlação

## Procedimentos gerais para a construção de um teste de hipótese para $\rho$

- Usar as hipóteses:

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

- Definir um nível de **significância**\*  $\alpha$  (ex.:  $\alpha 0,05$ ), que irá determinar o nível de **confiança**  $100(1 - \alpha)\%$  do teste
- Determinar a **região de rejeição** com base no nível de significância  $\rightarrow t_{crit}$  (com  $n - 2$  graus de liberdade)
- Calcular a **estatística de teste**, sob a hipótese nula

$$t_{calc} = r \sqrt{\frac{n - 2}{1 - r^2}}$$

- Rejeitar a hipótese nula se a estatística de teste calculada estiver dentro da região de rejeição ( $|t_{calc}| > |t_{crit}|$ )

# Exemplo

Usando os dados dos exemplos anteriores, realize os testes de hipótese para o coeficiente de correlação  $\rho$ , usando um nível de 5% de significância.

- Distância (km) x Consumo (l)  $\rightarrow r = 0.9475761$
- Peso (t) x Consumo (km/l)  $\rightarrow r = -0.8511547$

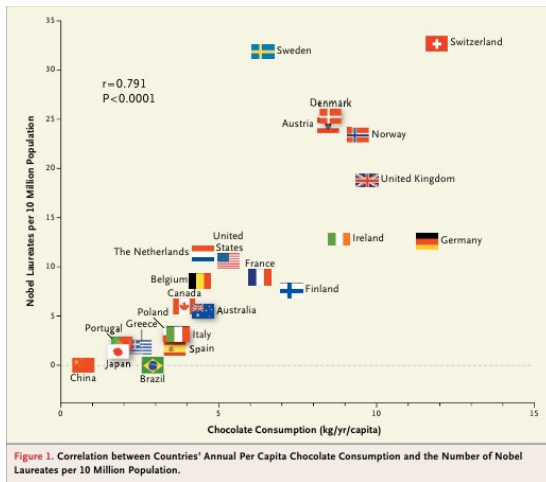
# Correlação

**ATENÇÃO!** — Correlação não implica causalção!

Existir uma correlação (positiva ou negativa) entre duas VAs  $X$  e  $Y$ , mesmo que significativa, **não** implica que  $X$  **causa**  $Y$ .



# Correlação



# Correlação

