# Examinando Relações nos Dados

## Mineração de Dados

Ronaldo C. Prati[1]

[1]Universidade Federal do ABC (UFABC), ronaldo.prati@ufabc.edu.br

# Introdução

# Overview

- Nas análise descritiva que fizemos, nos concentramos em uma variável
- Em muitas situações, é interessante examinar relações de duas (ou mais) variáveis por vez

# Variáveis explicatórias e resposta

Tipicamente, estamos interessados no relacionamento de uma variável explanatória e uma variável resposta.

- a **variável explanatória** (também chamada de variável independente) é aquela que usamos pra explicar, predizer ou supomos afetar a variável resposta
- a **variável resposta** (também chamada de variável dependente)

# Exemplos:

- Existe alguma relação entre renda e a nota em um exame padronizado?
- Como o número de calorias em um sanduíche é afetado pelo tipo de recheio?
- Os hábitos de fumo estão relacionados a idade?

# Variáveis explicatórias e resposta

- **Pergunta:** O papel dessas variávies é sempre claro? Em outras palavras, é sempre claro qual variável é a explantória e qual é a resposta?
- **Resposta:** Não. Existem situações em que essa classificaçõe não é clara. Há situações em que ambas as variáveis podem assimir esse papel (por exemplo, cosidere as notas entre os eixos do enem)

# Classificação quanto ao tipo

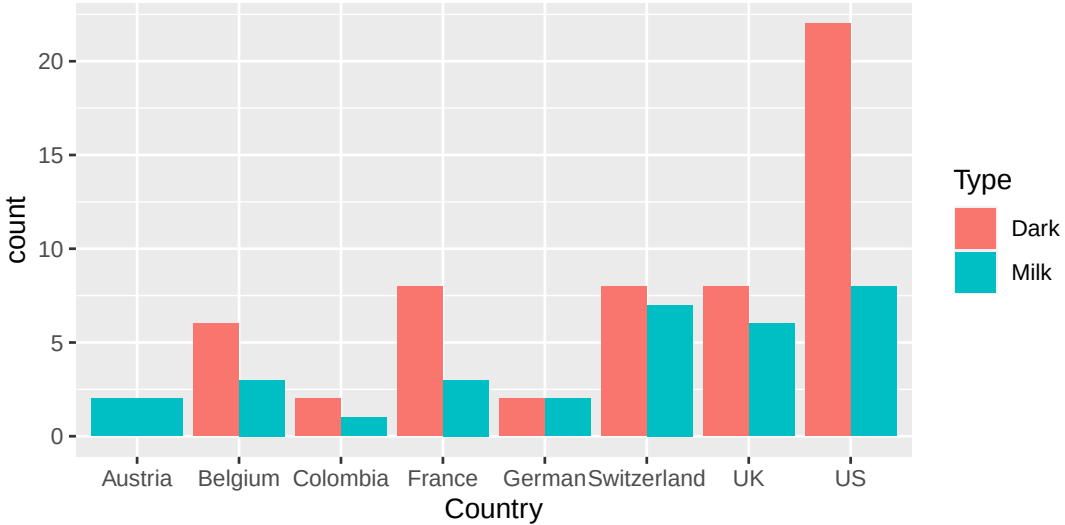
- Variável explanatória e resposta numéricas
- Variável explanatória e resposta categóricas
- Variável explanatória categórica e resposta numérica
- Variável explanatória numérica e resposta categórica

# Categórico x Categórico

# Tabelas de contingência

| | Austria | Belgium | Colombia | France | German | Switzerland | UK | US | Sum |
|---|---|---|---|---|---|---|---|---|---|
| Dark | 0 | 6 | 2 | 8 | 2 | 8 | 8 | 22 | 56 |
| Milk | 2 | 3 | 1 | 3 | 2 | 7 | 6 | 8 | 32 |
| Sum | 2 | 9 | 3 | 11 | 4 | 15 | 14 | 30 | 88 |

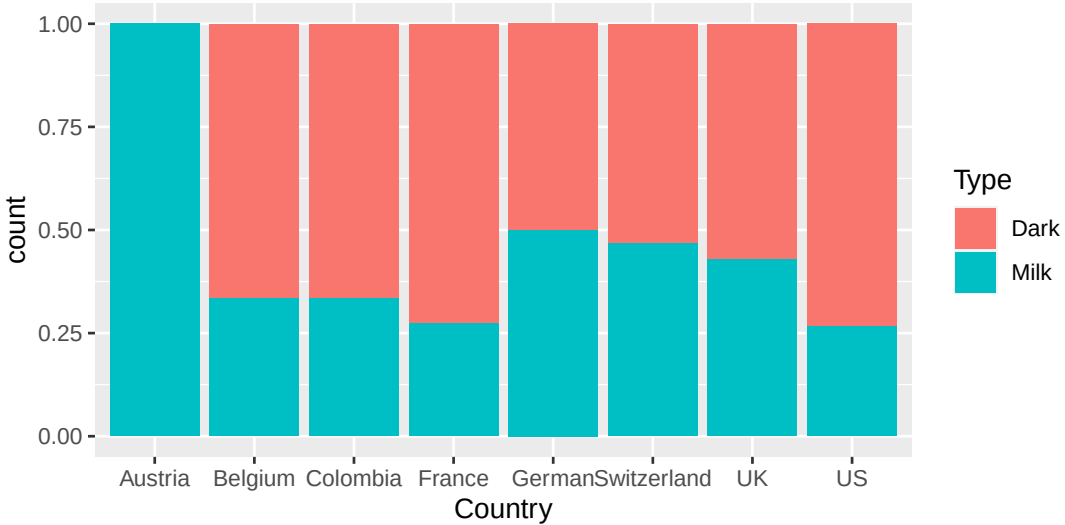
# Tabelas de contingência



# Tabelas de contingência

| | Austria | Belgium | Colombia | France | German | Switzerland | UK | US | Sum |
|---|---|---|---|---|---|---|---|---|---|
| Dark | 0.0 | 6.8 | 2.3 | 9.1 | 2.3 | 9.1 | 9.1 | 25.0 | 63.6 |
| Milk | 2.3 | 3.4 | 1.1 | 3.4 | 2.3 | 8.0 | 6.8 | 9.1 | 36.4 |

# Tabelas de contingência



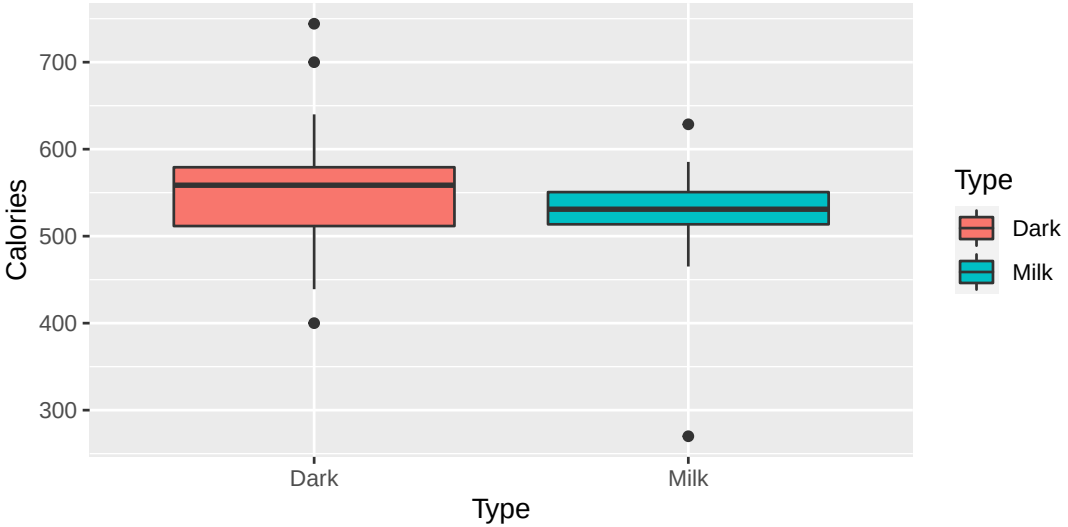
# Categórico x Contínuo

# Estatística por grupo

- Variável explicativa -> Tipo (categórica)
- Variável resposta -> Calorias (contínua)

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|---|
| Dark | 400 | 511.6279 | 558.5698 | 550.8751 | 579.2105 | 744.1860 |
| Milk | 270 | 513.6213 | 530.9762 | 526.9849 | 550.6579 | 628.5714 |

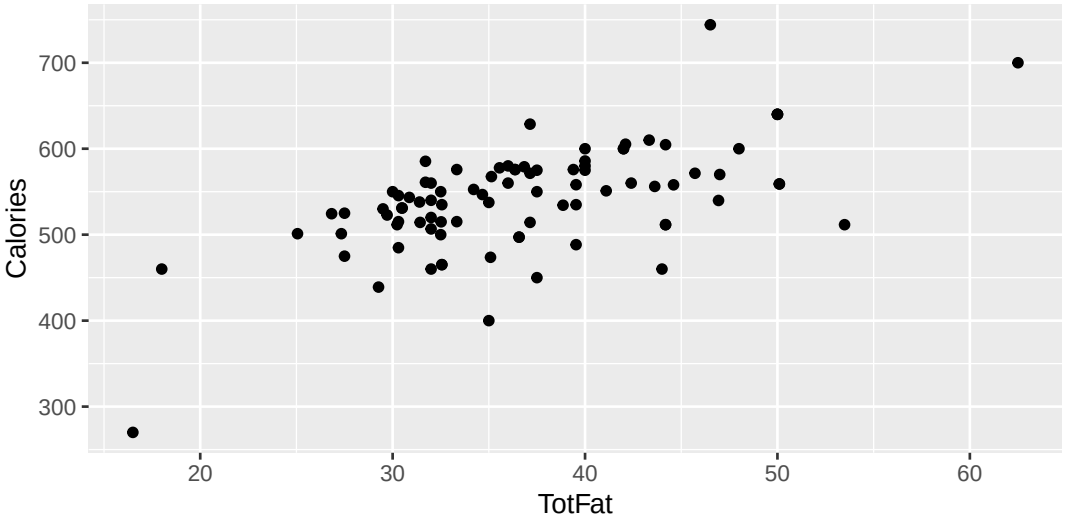
# Estatística por grupo (boxplot)



# Contínuo x Contínuo

# Gráfico de Dispersão

- Variável explicativa -> Tipo (categórica)
- Variável resposta -> Calorias (contínua)

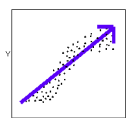
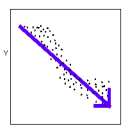
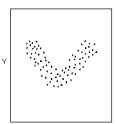


# Gráfico de Dispersão

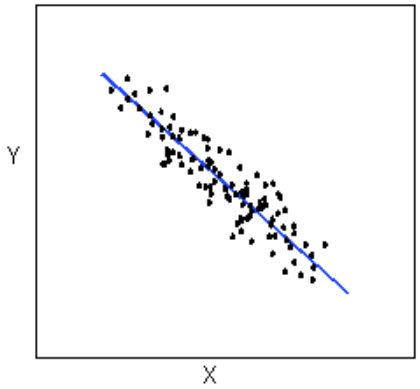
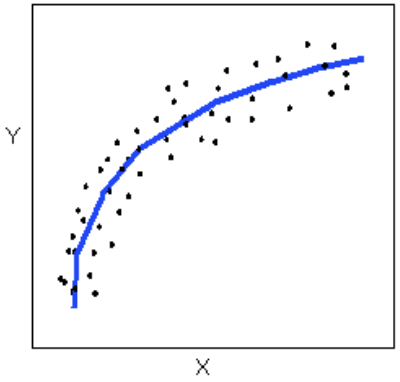
Quando analisamos relações entre duas variáveis com gráficos de dispersão, procuramos por

- Padrão geral:
  - Direção
  - Forma
  - "Força"
- Desvio do padrão
  - Valores discrepantes (outliers)

# Gráfico de Dispersão

| Relação Positiva | Relação Negativa | Nem positiva nem gegativa |
|---|---|---|
|  |  |  |

# Gráfico de Dispersão

| Relação Linear | Relação CurveLinear |
| --- | --- |
|  |  |



- Existem outros possíveis formatos, mas esses são os mais comuns e fáceis de identificar

# Gráfico de Dispersão

Outro padrão comum é a formação de grupos

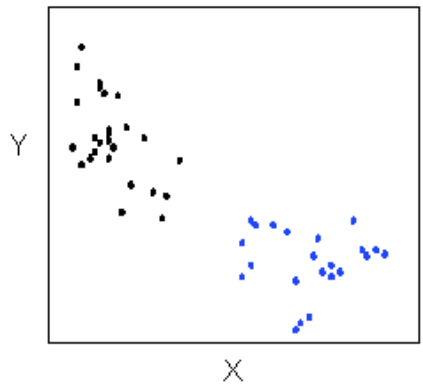
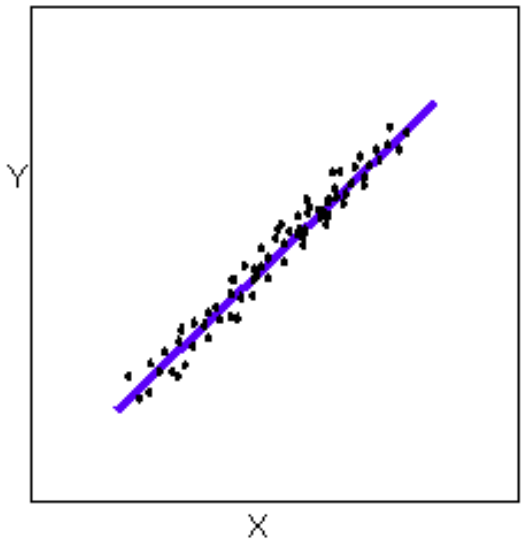


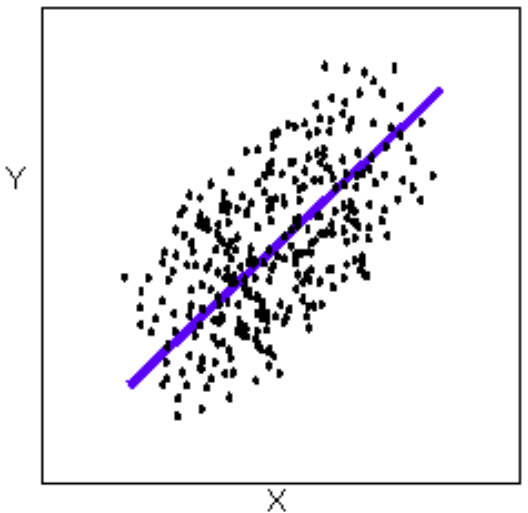

Figure 1: image

# Gráfico de Dispersão

Relação Forte




Relação Fraca



# Gráfico de Dispersão

Outro padrão comum é a formação de grupos

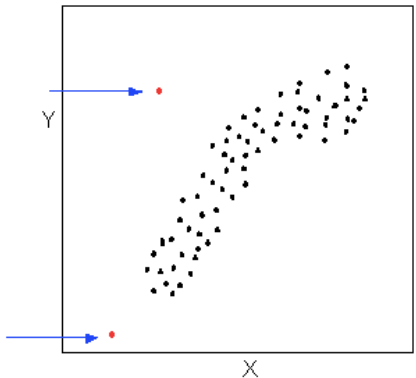
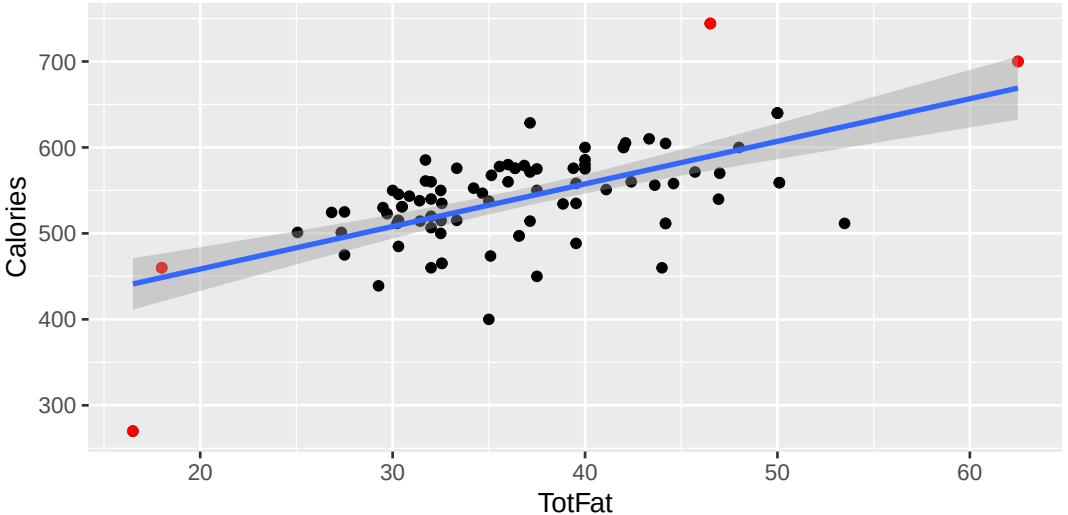


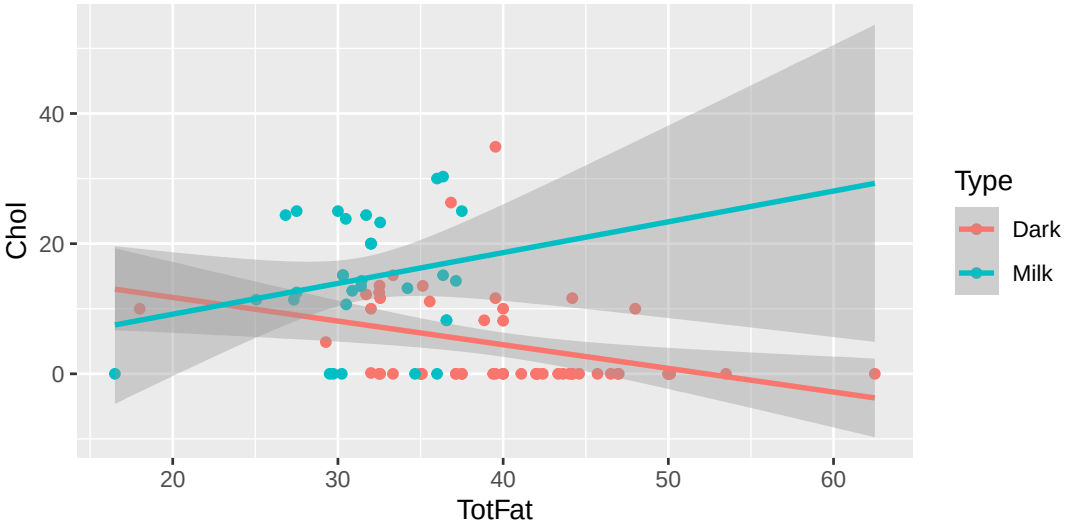

Figure 2: image

# Exemplo

Vamos analisar a relação entre o total de gordura (TotFat) e calories (Calories) da base de dados Chocolate



# Gráfico de Dispersão Rotulado



# Relação Linear

- Uma inspeção visual de gráficos de dispersão nos ajuda a analisar a direção, forma e força da relação entre duas variáveis
- Entretanto, essa inspeção é subjetiva, e seria interessante usar um método mais objetivo de mensurar essa relação

# Relação Linear

- Vamos nos focar no caso especial de **relações que tem uma forma linear**, uma vez que elas são comuns e fáceis de identificar
- É importante ressaltar que **nem toda a relação entre duas variávies numéricas é tem uma forma linear**
- O método que veremos é **apropriado apenas para exeminar relações lineares**

# Coeficiente de Correlação

- Uma medida numérica que avalia a força da relação linear é chamada de **coeficiente de correlação**
- **Definição:** O **coeficiente de correlação** ($r$) é uma medida numérica da **força** e **direção** da relaão linear entre duas variáveis quantitativas.

**Cálculo:** $r$ é calculado com a fórmula:

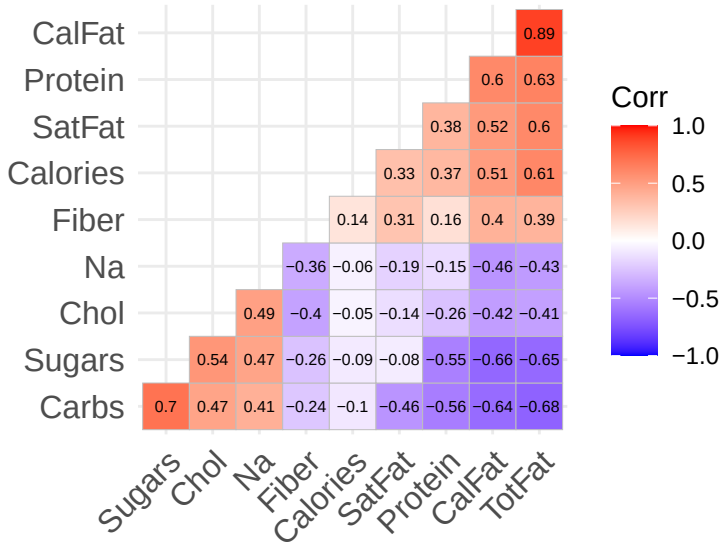
$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{S_x} \right) \left( \frac{y_i - \bar{y}}{S_y} \right)$$

em que $S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$ (analogamente para $S_y$) é o desvio padrão amostral. Muitos pacotes (incluindo em Python) implementam o cálculo do coeficiente de correlação.

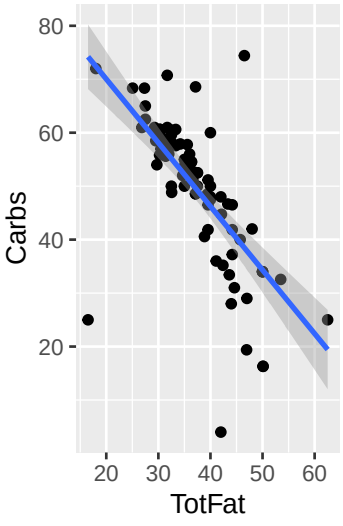
# Matriz de correlação

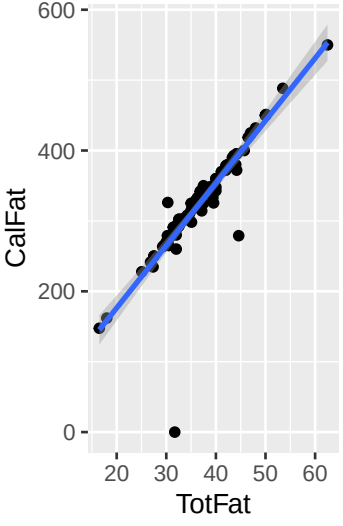
| | Calories | CalFat | TotFat | SatFat | Chol | Na | Carbs | Fiber | Sugars | Prot |
|---|---|---|---|---|---|---|---|---|---|---|
| Calories | 1.00 | 0.51 | 0.61 | 0.33 | -0.05 | -0.06 | -0.10 | 0.14 | -0.09 | 0 |
| CalFat | 0.51 | 1.00 | 0.89 | 0.52 | -0.42 | -0.46 | -0.64 | 0.40 | -0.66 | 0 |
| TotFat | 0.61 | 0.89 | 1.00 | 0.60 | -0.41 | -0.43 | -0.68 | 0.39 | -0.65 | 0 |
| SatFat | 0.33 | 0.52 | 0.60 | 1.00 | -0.14 | -0.19 | -0.46 | 0.31 | -0.08 | 0 |
| Chol | -0.05 | -0.42 | -0.41 | -0.14 | 1.00 | 0.49 | 0.47 | -0.40 | 0.54 | -0 |
| Na | -0.06 | -0.46 | -0.43 | -0.19 | 0.49 | 1.00 | 0.41 | -0.36 | 0.47 | -0 |
| Carbs | -0.10 | -0.64 | -0.68 | -0.46 | 0.47 | 0.41 | 1.00 | -0.24 | 0.70 | -0 |
| Fiber | 0.14 | 0.40 | 0.39 | 0.31 | -0.40 | -0.36 | -0.24 | 1.00 | -0.26 | 0 |
| Sugars | -0.09 | -0.66 | -0.65 | -0.08 | 0.54 | 0.47 | 0.70 | -0.26 | 1.00 | -0 |
| Protein | 0.37 | 0.60 | 0.63 | 0.38 | -0.26 | -0.15 | -0.56 | 0.16 | -0.55 | 1 |

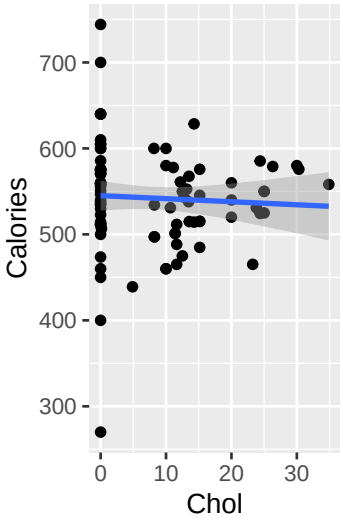
# Matriz de correlação



# Matriz de correlação

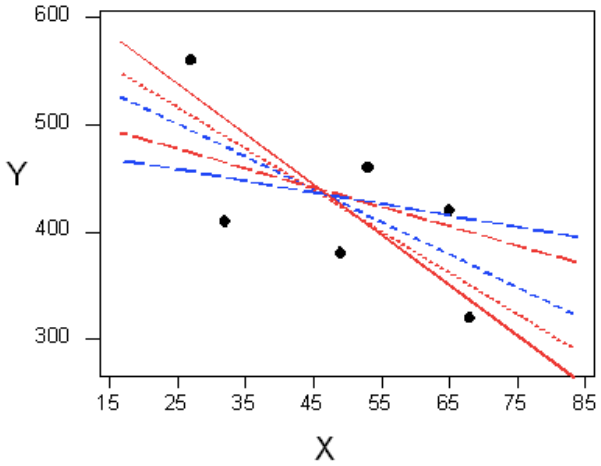


# Regressão Linear

Existem muitas retas que podemos usar como a que melhor se adequa à tendencia dos dados?



# Regressão Linear

Nos queremos encontrar a reta que minimiza a distância média para os dados:

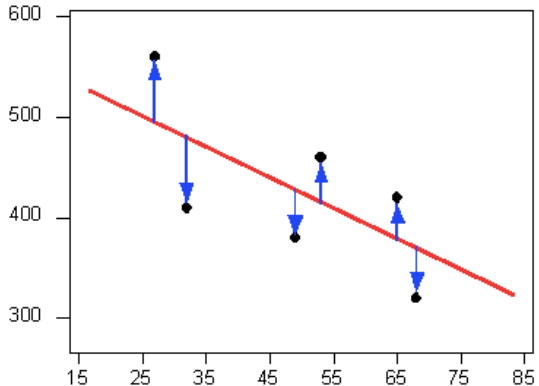



Figure 4: regressao

# Mínimos quadrados

Uma estratégia para encontrar essa reta é buscar pela reta que minimiza o quadrado das distâncias. Essa abordagem é conhecida como mínimos quadrados.

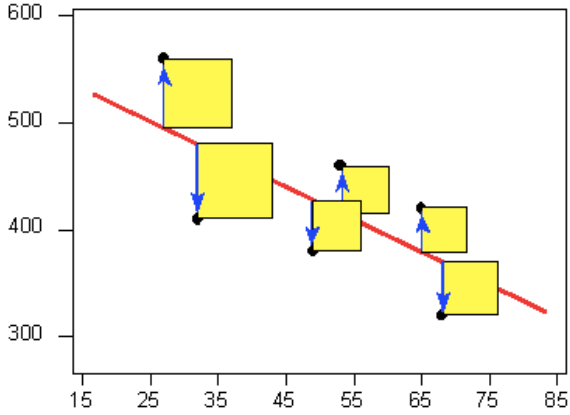



Figure 5: regressao

# Equação da reta

Considere a equação da reta:

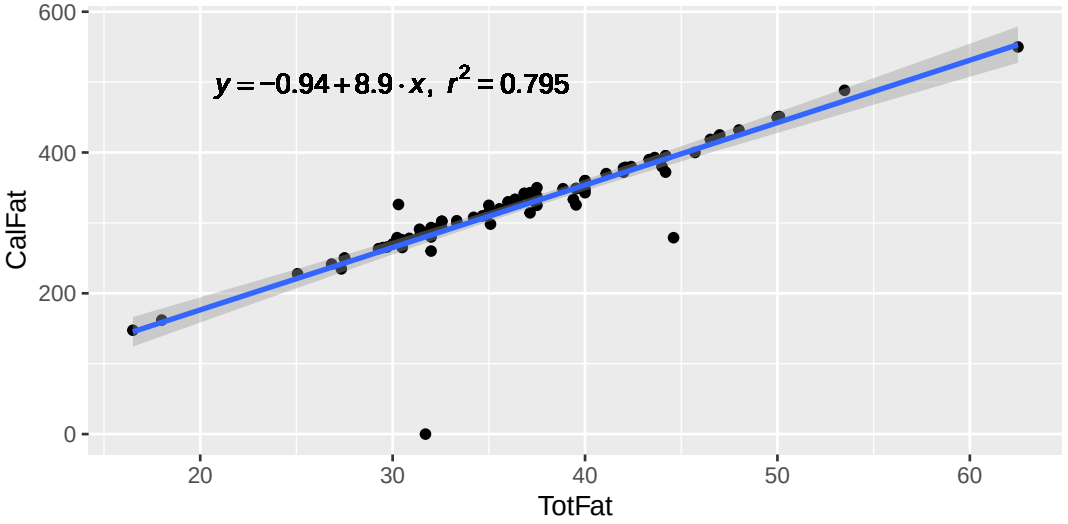
$$Y = aX + b$$

Podemos obter os parâmetros $a$ e $b$ que minimizam o quadrado das distâncias

$$b = r\left(\frac{S_x}{S_y}\right)$$

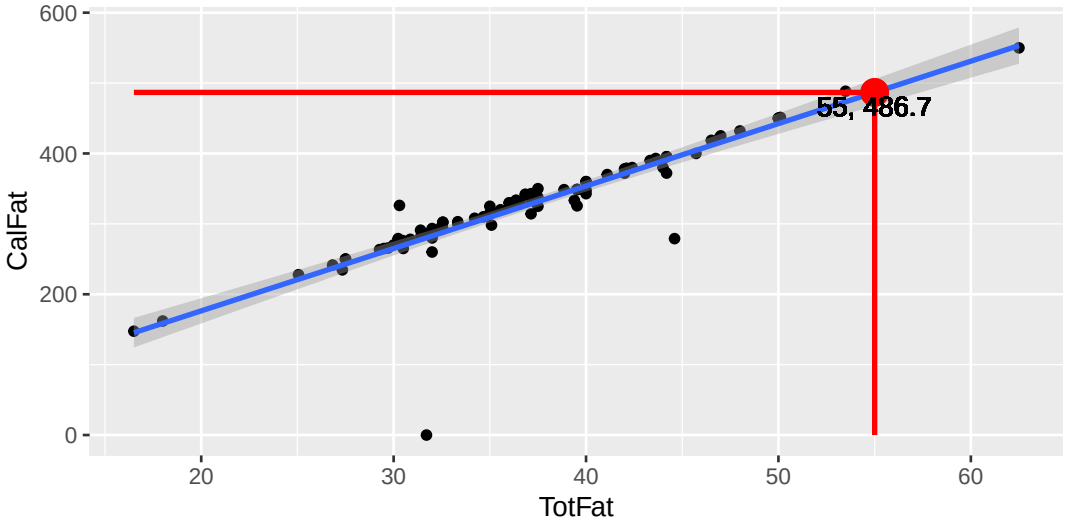
$$a = \bar{Y} - b\bar{X}$$

# Equação da reta



# Predizendo um valor

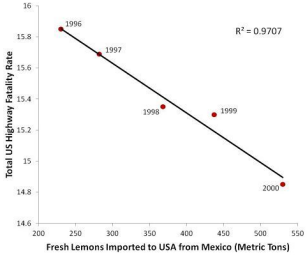
Essa equação pode ser usada para predizer um valor, como mostrado a seguir

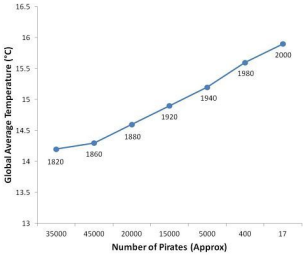


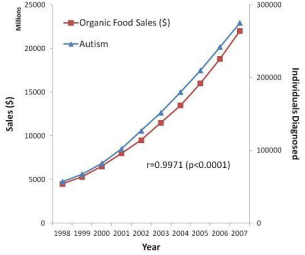
# Associação X Causa

Uma associação entre duas variáveis não é evidência suficiente para justificar uma relação causal entre elas

| Limão X Acidentes | Piratas X Aquecimento | Comida Orgânica X Autismo |
|---|---|---|



# Paradoxo de Simpson

Absoluto

| | | Patient's Status | | |
|---|---|---|---|---|
| | | Died | Survived | **Total** |
| Hospital | Hospital A | 63 | 2037 | 2100 |
| | Hospital B | 16 | 784 | 800 |
| | **Total** | 79 | 2821 | 2900 |

Relativo

| | | Patient's Status | | |
|---|---|---|---|---|
| | | Died | Survived | **Total** |
| Hospital | Hospital A | **3%** | 97% | 100% |
| | Hospital B | **2%** | 98% | 100% |

# Paradoxo de Simpson

- Aparentemente, o Hospital A tem uma mortalidade 50% maior que o hospital B.
- Essa constatação está correta?
- E se o Hospital A receber casos mais graves que o hospital B? Nesse caso, devemos levar em consideração a variável oculta severidade da doença

# Paradoxo de Simpson

| | | Patient's Status | | |
|---|---|---|---|---|
| | | Died | Survived | **Total** |
| Hospital | Hospital A | 63 | 2037 | 2100 |
| | Hospital B | 16 | 784 | 800 |
| | **Total** | 79 | 2821 | 2900 |

Accounting for the lurking variable: **"severity of illness"**

**Patients severly ill**

| | | Patient's Status | | |
|---|---|---|---|---|
| | | Died | Survived | **Total** |
| Hospital | Hospital A | 57 | 1443 | 1500 |
| | Hospital B | 8 | 192 | 200 |
| | **Total** | 65 | 1635 | 1700 |

**Patients *not* severly ill**

| | | Patient's Status | | |
|---|---|---|---|---|
| | | Died | Survived | **Total** |
| Hospital | Hospital A | 6 | 594 | 600 |
| | Hospital B | 8 | 592 | 600 |
| | **Total** | 14 | 1186 | 1200 |

# Paradoxo de Simpson

**Patients severly ill**

| | | Patient's Status | | |
|---|---|---|---|---|
| | | Died | Survived | **Total** |
| Hospital | Hospital A | **3.8%** | 96.2% | 100% |
| | Hospital B | **4.0%** | 96.0% | 100% |

**Patients *not* severly ill**

| | | Patient's Status | | |
|---|---|---|---|---|
| | | Died | Survived | **Total** |
| Hospital | Hospital A | **1.0%** | 99.0% | 100% |
| | Hospital B | **1.3%** | 98.7% | 100% |

# Paradoxo de Simpson

- O uso da variável oculta pode pode levar a uma alteração no sentido da associação!
- Quando a inclusão de uma variável nos leva a repensar a direção da associação, isso é conhecido como **Paradoxo de Simpson**