

Análise Quantitativa de Dados em Linguística

Régressão linear simples

Ronaldo Lima Jr.

ronaldojr@letras.ufc.br

ronaldolimajr.github.io

Universidade Federal do Ceará

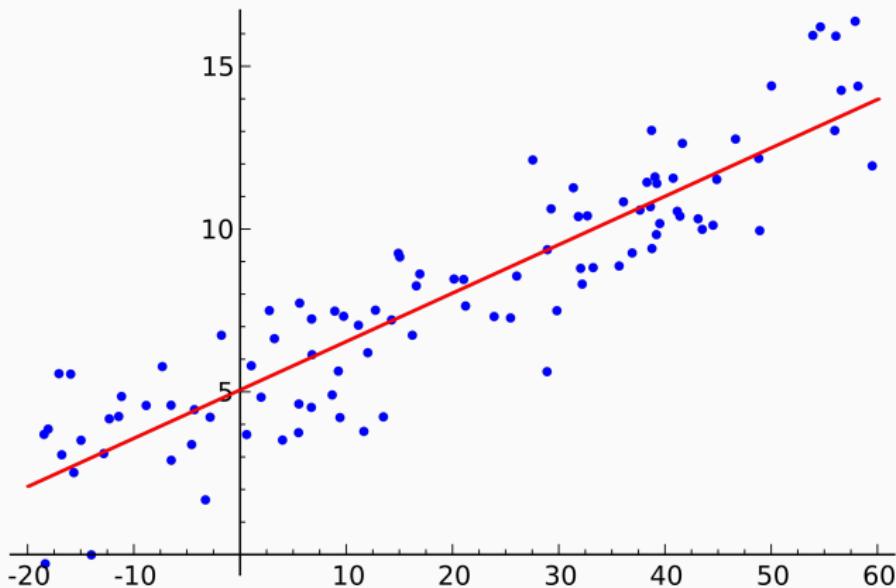
Roteiro

1. Correlação
2. Regressão Linear Simples
3. Modelo com previsor categórico

Correlação

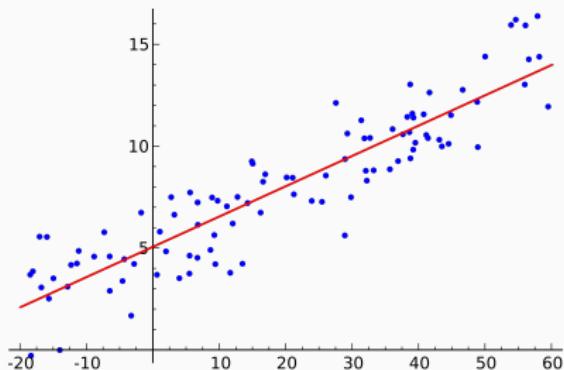
Correlação

- Precisamos partir de uma correlação



Correlação

Precisamos partir de uma correlação

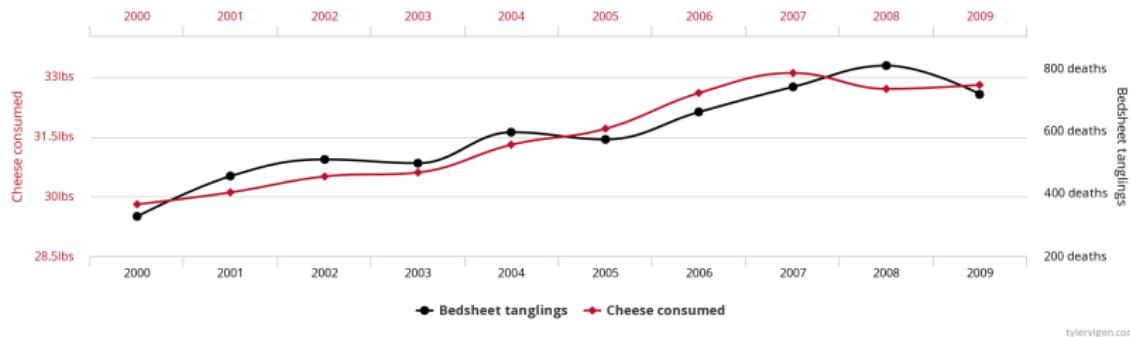


- Porém, **correlação** não é sinônimo de **motivação**

Correlation does not mean causation

Correlação

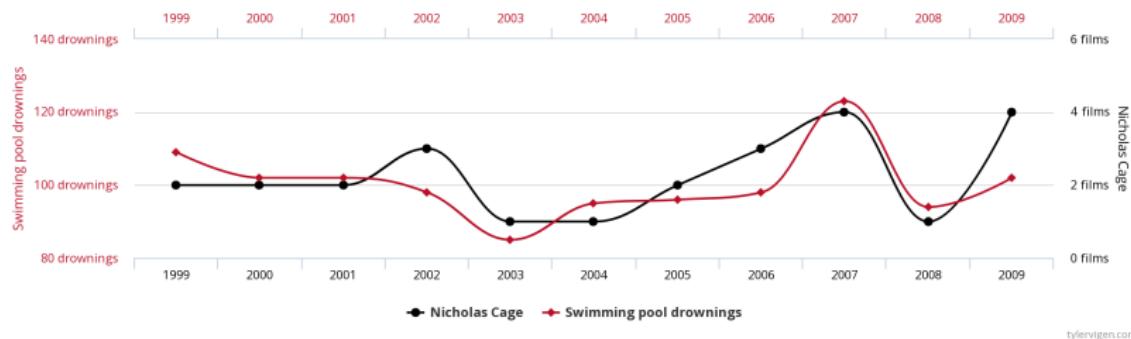
Per capita cheese consumption
correlates with
Number of people who died by becoming tangled in their bedsheets



<http://www.tylervigen.com/spurious-correlations>

Correlação

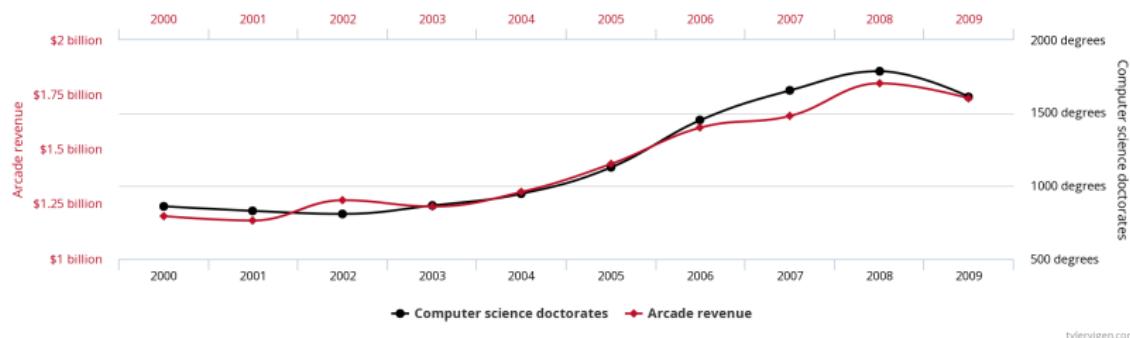
Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in



<http://www.tylervigen.com/spurious-correlations>

Correlação

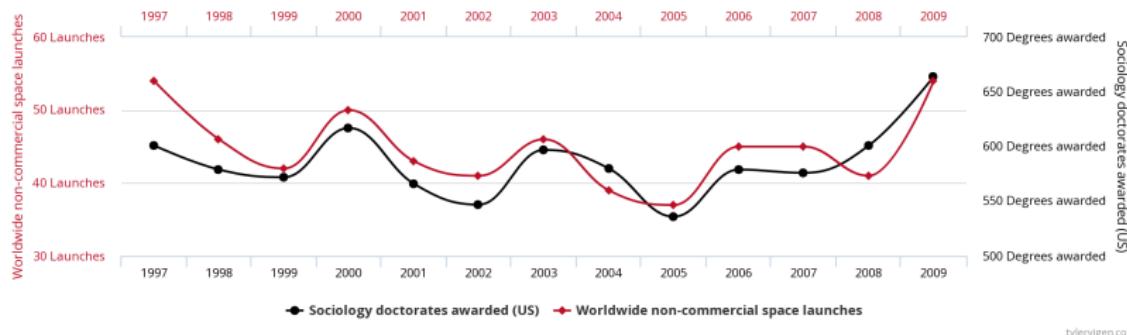
Total revenue generated by arcades
correlates with
Computer science doctorates awarded in the US



<http://www.tylervigen.com/spurious-correlations>

Correlação

Worldwide non-commercial space launches correlates with Sociology doctorates awarded (US)



<http://www.tylervigen.com/spurious-correlations>

Correlação

Repita comigo:

- **correlação** não é sinônimo de **motivação**

One more time:

- **Correlation does not mean causation**

Contudo, só haverá relação de efeito (*causation*) se houver correlação

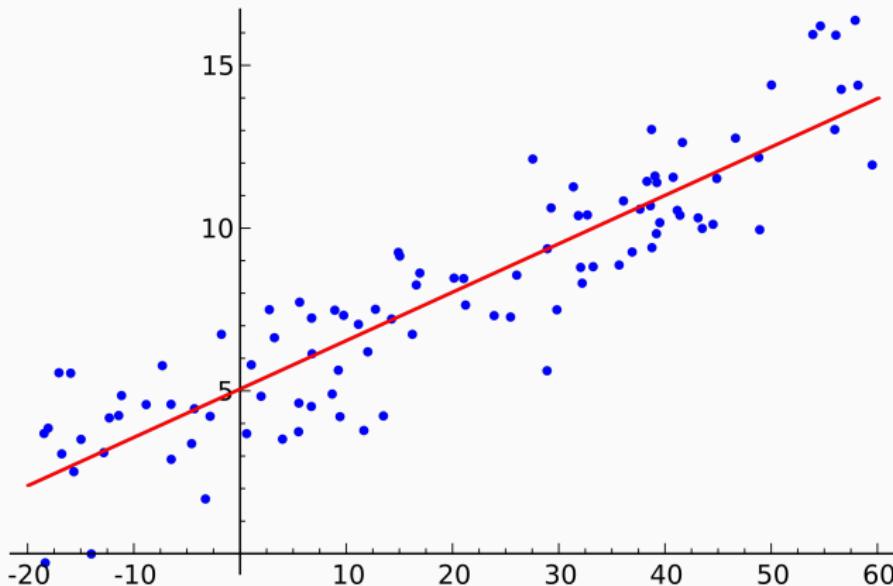
- Correlação é requisito para a regressão

Regressão Linear Simples

Regressão Linear

Objetivos:

1. Verificar se há uma relação entre variáveis
2. Prever valores não observados



Regressão Linear

- Um Modelo de Regressão Linear estima o valor de y baseado no valor de x

y = variável dependente = variável resposta (*response variable*)

x = variável independente = variável preditora/previsora
(*predictor/explanatory variable*)

- A variável dependente/resposta deve ser numérica (contínua)
(para variável resposta categórica utiliza-se regressão logística)
- A variável independente/preditora pode ser numérica ou categórica e pode ser mais de uma (regressão múltipla/multifatorial)

Regressão Linear

Em um teste de correlação não importa a ordem das variáveis

Em um modelo linear importa, pois estimamos o valor de y em função de x (i.e., dado um valor de x)

Portanto, a ordem deve ser $y \sim x$

- Equação linear:

$$y = a + bx$$

$$y = \beta_0 + \beta_1 x$$

y valor que será estimado (e.g. altura)

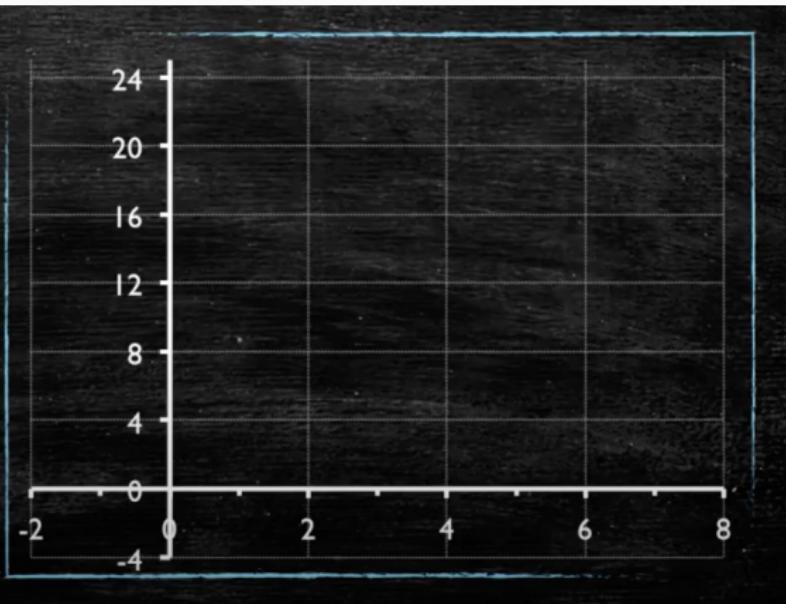
x valor que será dado (e.g. idade)

β_0 e β_1 valores fornecidos pelo modelo (coeficientes)

- Equação → dados coletados → modelo → dados estimados

Regressão Linear

$$y = \beta_0 + \beta_1 x$$

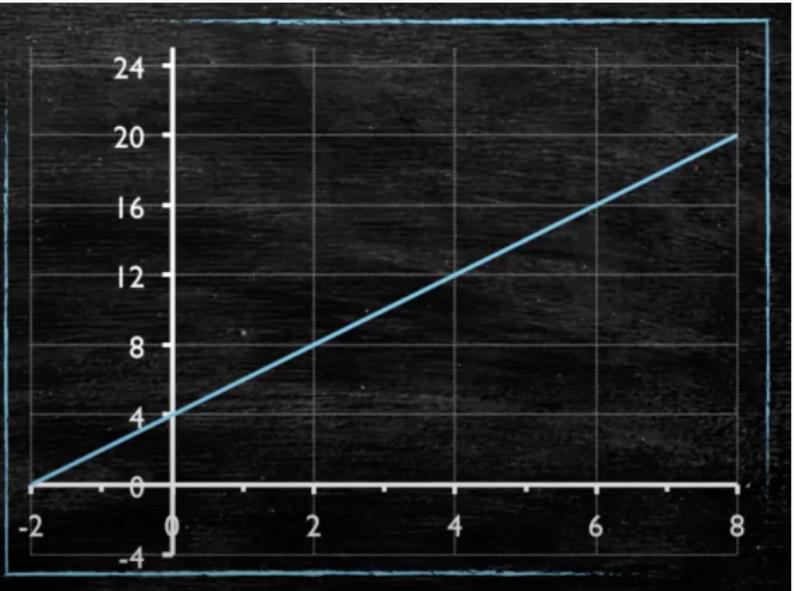


<https://www.youtube.com/watch?v=owI7zxCqNY0>

Regressão Linear

$$y = \beta_0 + \beta_1 x$$

$$y = 4 + 2x$$



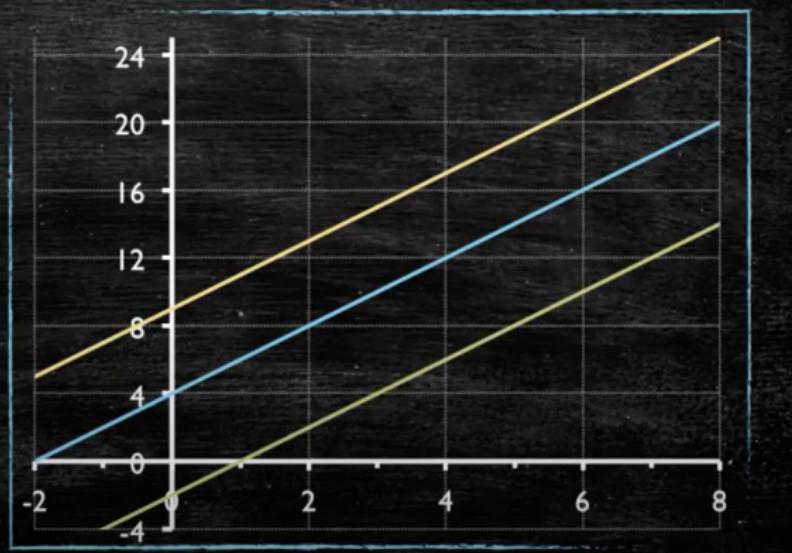
<https://www.youtube.com/watch?v=owI7zxCqNY0>

Regressão Linear

$$y = 4 + 2x$$

$$y = 9 + 2x$$

$$y = -2 + 2x$$



<https://www.youtube.com/watch?v=owI7zxCqNY0>

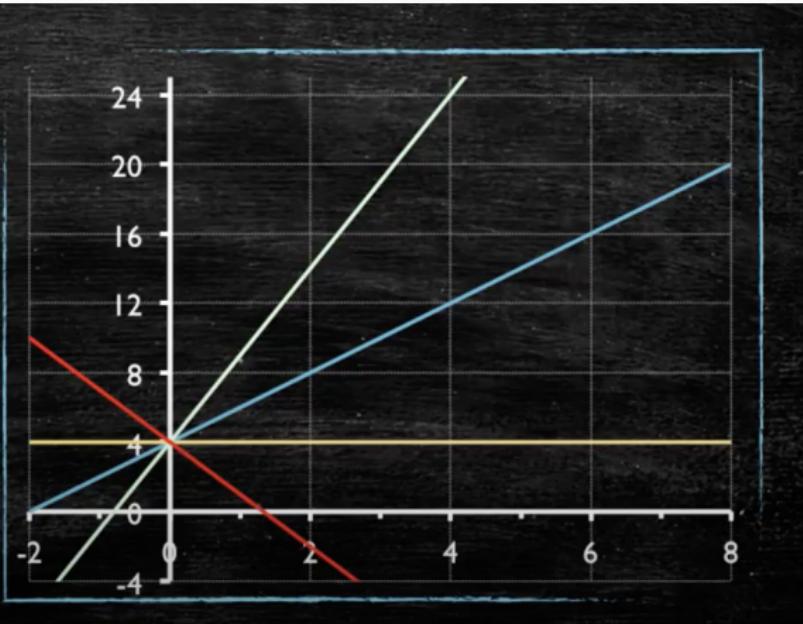
Regressão Linear

$$y = 4 + 2x$$

$$y = 4 + 5x$$

$$\begin{aligned}y &= 4 + 0x \\&= 4\end{aligned}$$

$$y = 4 - 3x$$

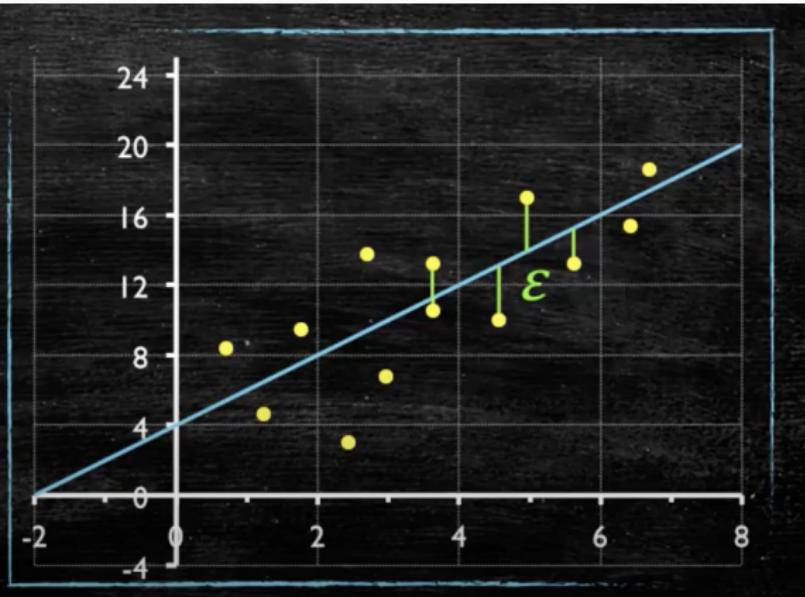


<https://www.youtube.com/watch?v=owI7zxCqNY0>

Regressão Linear

$$y = 4 + 2x$$

True Values



- valor observado – valor estimado = resíduo
- a reta de regressão faz um caminho de maneira a gerar os menores resíduos possíveis

Exemplo 1

Modelo de regressão para estimar a altura em função da idade
(dada uma idade)

```
1 modAltura = lm(altura ~ idade, data = cor)
2 modAltura
3 
3 Coefficients:
4 (Intercept)      idade
5       62.504        7.545
```

- São apresentados dois coeficientes: o *intercept* (coeficiente linear) e o *slope* (coeficiente angular)
- O *intercept* indica onde a reta de regressão passa no y quando $x = 0$
- O *slope* diz quanto há de aumento (ou diminuição) de y em cada unidade de aumento de x

Exemplo 1

Modelo de regressão para estimar a altura em função da idade
(dada uma idade)

```
1 modAltura = lm(altura ~ idade, data = cor)
2 modAltura
3 
3 Coefficients:
4 (Intercept)      idade
5       62.504        7.545
```

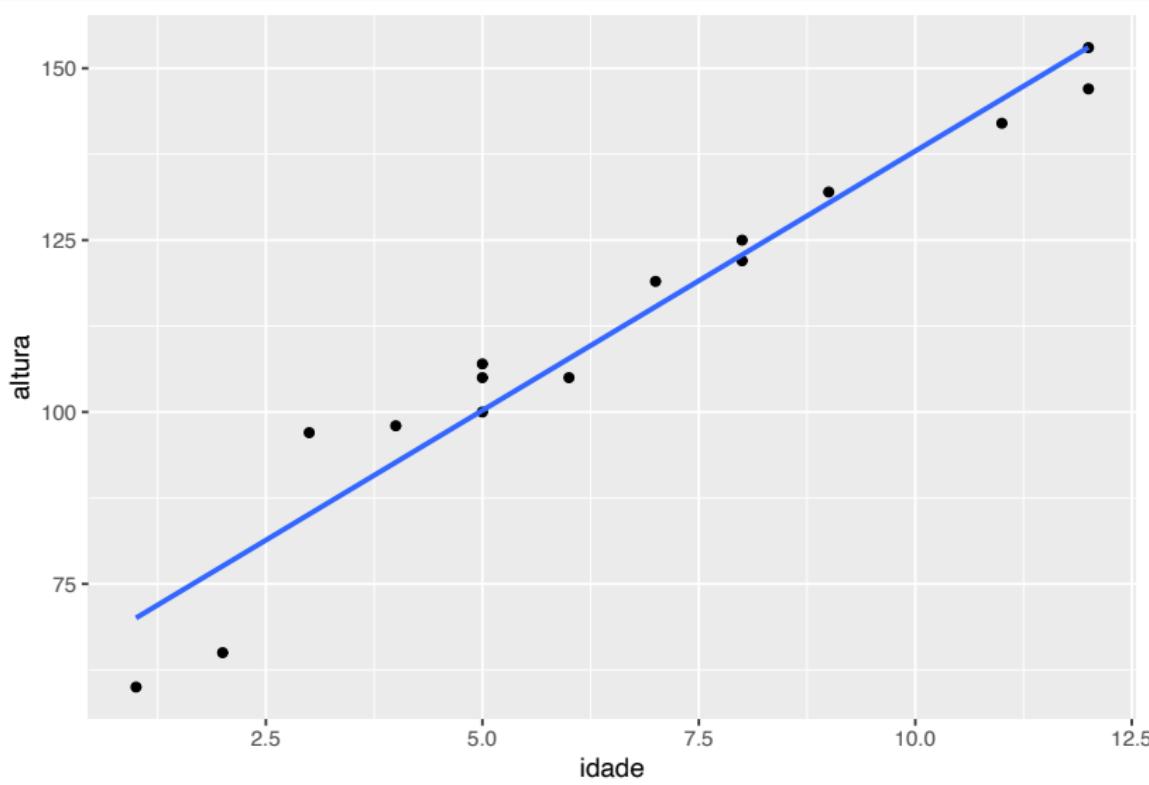
- Este modelo estima 62,5cm de altura quando idade = 0 e um aumento de 7,5cm para cada ano a mais de idade

$$\beta_0 = 62,5$$

$$\beta_1 = 7,5 \rightarrow \text{tamanho do efeito}$$

Exemplo 1

$$\beta_0 = 62,5 \text{ e } \beta_1 = 7,5$$



Exemplo 1

$$\beta_0 = 62,5 \text{ e } \beta_1 = 7,5$$

- Não são valores presentes nos dados, nem valores do mundo real, são valores do **modelo**

"All models are wrong, but some are useful" George E. P. Box

"The Golem of Prague" Richard McElreath

- Os dados que alimentam o modelo (*fit*) farão dele um modelo de estimativa melhor ou pior
- Nosso modelo estima um recém-nascido de 62,5cm
 - raramente o intercept é muito informativo
- Nosso modelo estima um aumento de 7,5cm para cada ano que aumenta na idade
 - o slope é bastante informativo, é o tamanho do efeito (*effect size*)

Cuidado! O modelo é um robô:

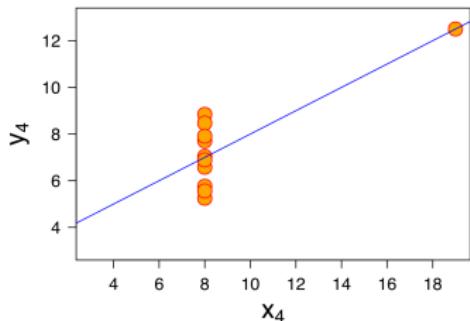
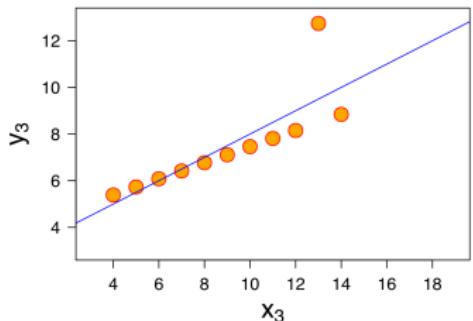
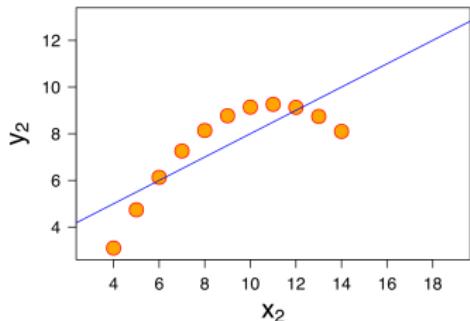
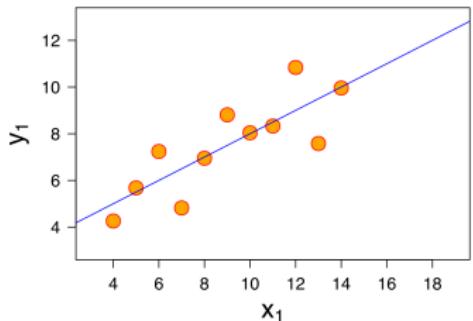


Figure 1: Quarteto de Anscombe – mesma linha de regressão, mesma média, mesmo desvio-padrão e mesmas correlações

Cuidado! O modelo é um robô:

- No nosso modelo altura \sim idade, qual seria a altura de alguém de 12 anos?

$$y = \beta_0 + \beta_1 x$$

$$\text{altura} = 62,5 + 7,5 \times 12$$

$$\rightarrow \text{altura} = \textcolor{orange}{152.5\text{cm}}$$

- E a altura de alguém de 60 anos?

$$y = \beta_0 + \beta_1 x$$

$$\text{altura} = 62,5 + 7,5 \times 60$$

$$\rightarrow \text{altura} = \textcolor{orange}{512.5\text{cm}} (!)$$

Cuidado! O modelo é um robô:

- E a altura de alguém de 60 anos?

$$y = \beta_0 + \beta_1 x$$

$$\text{altura} = 62,5 + 7,5 \times 60$$

$$\rightarrow \text{altura} = \textcolor{orange}{512.5\text{cm}} \text{ (!)}$$

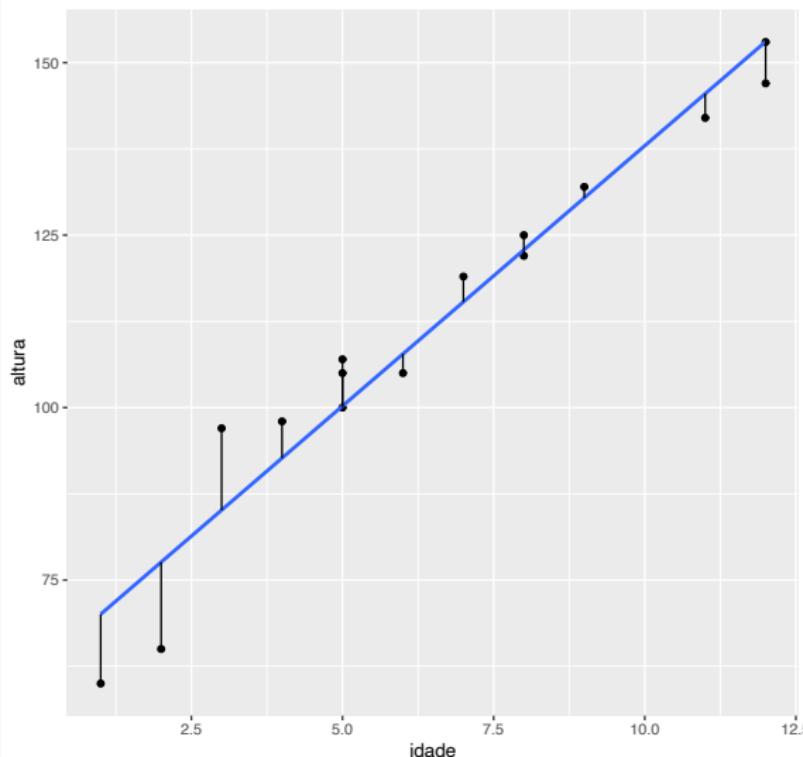
- Nosso modelo não serve para adultos
só sabemos disso porque **conhecemos** essa realidade
mas normalmente queremos **inferir** uma realidade desconhecida a partir de observações

Resíduos

```
1 # Dados de altura observados:  
2 cor$altura  
  
3 60 65 97 98 100 105 107 105 119 122 125 132 142 147 153  
  
4 # Dados estimados pelo modelo:  
5 modAltura$fitted.values  
  
6      1          2          3          4          5  
7 70.04928 77.59459 85.13990 92.68521 100.23052 (...)  
  
8 # Resíduos (diferença entre valores observados e valores estimados)  
9 modAltura$residuals  
  
10     1          2          3          4  
11 -10.04928458 -12.59459459 11.86009539  5.31478537 (...)
```

Resíduos

```
1 | # Resíduos:  
2 |   1           2           3           4  
3 | -10.04928458 -12.59459459 11.86009539 5.31478537 (...)
```



Informações completas do modelo

```
1 summary(modAltura)

2 Call:
3 lm(formula = altura ~ idade, data = cor)

4 Residuals:
5      Min       1Q   Median       3Q      Max
6 -12.5946  -3.1391  -0.0477  4.2242  11.8601

7 Coefficients:
8             Estimate Std. Error t value Pr(>|t|)
9 (Intercept) 62.5040    3.7691   16.58 3.98e-10 ***
10 idade        7.5453    0.5135   14.69 1.78e-09 ***
11 ---
12 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

13 Residual standard error: 6.651 on 13 degrees of freedom
14 Multiple R-squared:  0.9432,          Adjusted R-squared:  0.9388
15 F-statistic: 215.9 on 1 and 13 DF,  p-value: 1.782e-09
```

Informações completas do modelo

```
1 | Call:  
2 | lm(formula = altura ~ idade, data = cor)
```

→ fórmula que utilizamos

```
1 | Residuals:  
2 |    Min      1Q   Median      3Q      Max  
3 | -12.5946 -3.1391 -0.0477  4.2242 11.8601
```

→ distribuição dos resíduos em mínimo, 1º quartil, mediana, 3º quartil e máximo

- inspecione visualmente se os resíduos estão minimamente simétricos (mediana ≈ 0 , quartis relativamente espelhados, min e max também)

Se o modelo erra, melhor que erre tanto para mais como para menos

Informações completas do modelo

```
1 Coefficients:  
2             Estimate Std. Error t value Pr(>|t|)  
3 (Intercept) 62.5040     3.7691   16.58 3.98e-10 ***  
4 idade        7.5453     0.5135   14.69 1.78e-09 ***  
5 ---  
6 Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Coeficientes (*intercept* e *slope*)
- Erro padrão de cada coeficiente
- Teste-t para cada coeficiente (H_0 coeficiente = 0)
 - se o intercept não cruza o zero, $p < 0.05 \rightarrow$ **irrelevante**
 - se há ângulo significativo da reta de regressão, $p < 0.05 \rightarrow$ **há correlação**
é o teste-t que aparece no output de `cor.test()`
- Níveis de significância sugeridos pelo R

Informações completas do modelo

```
1 Residual standard error: 6.651 on 13 degrees of freedom
2 Multiple R-squared:  0.9432,      Adjusted R-squared:  0.9388
3 F-statistic: 215.9 on 1 and 13 DF,  p-value: 1.782e-09
```

- Erro padrão dos resíduos
- R^2 (r de Pearson ao quadrado) indica o quanto de variabilidade na variável resposta é explicada pelas variáveis incluídas no modelo
 - neste caso, 94% da altura pode ser explicada/prevista pela idade
- R^2 ajustado é um ajuste a depender do número de variáveis preditoras incluídas no modelo (em regressão múltipla)

Informações completas do modelo

```
1 | Residual standard error: 6.651 on 13 degrees of freedom
2 | Multiple R-squared:  0.9432,      Adjusted R-squared:  0.9388
3 | F-statistic: 215.9 on 1 and 13 DF,  p-value: 1.782e-09
```

→ Teste-F compara o modelo que criamos a um modelo sem variáveis preditoras (*intercept-only model*)

- testa a H_0 de que o modelo sem previsores é tão bom quanto o nosso modelo
- $p < 0.05$ indica que as variáveis preditoras melhoraram o modelo
- É utilizado para uma avaliação do modelo como um todo e para comparação de modelos

Intervalos de confiança dos coeficientes

$$\beta_0 = 62,5 \text{ e } \beta_1 = 7,5$$

```
1 | Calcular os intervalos de confiança (95%) dos coeficientes:  
2 | confint.lm(modAltura)  
3 |          2.5 %    97.5 %  
4 | (Intercept) 54.361405 70.646544  
5 | idade        6.435876  8.654744
```

Bônus: por que uma linha reta?

Porque pedimos para o robô (modelo) desenhar uma linha reta!

- E a altura de alguém de 60 anos?

$$y = \beta_0 + \beta_1 x$$

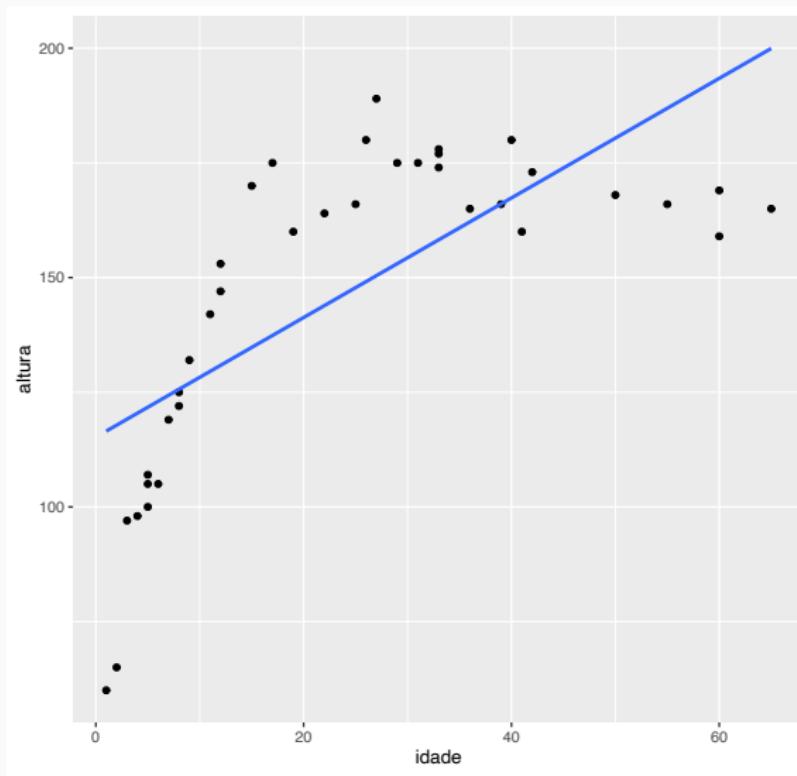
$$\text{altura} = 62,5 + 7,5 \times 60$$

$$\rightarrow \text{altura} = \textcolor{orange}{512.5\text{cm}} \text{ (!)}$$

E se acrescentarmos dados de adultos neste modelo?

Bônus: por que uma linha reta?

E se acrescentarmos dados de adultos?



$$\beta_0 = 115.2$$

$$\beta_1 = 1.30 \quad p < 0.001$$

$$R^2 = 0.5$$

$$F = 34.78 \quad p < 0.001$$

É um bom modelo
para prever altura?

Linhas curvas em regressão linear

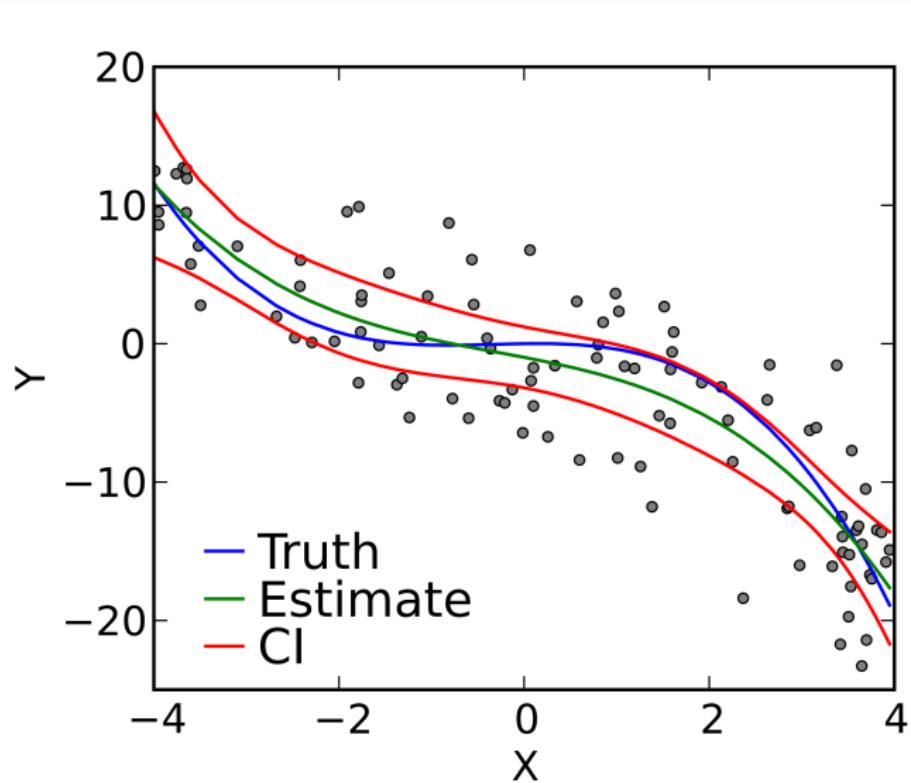


Figure 2: Cubic Polynomial Regression

Linhas curvas em regressão linear

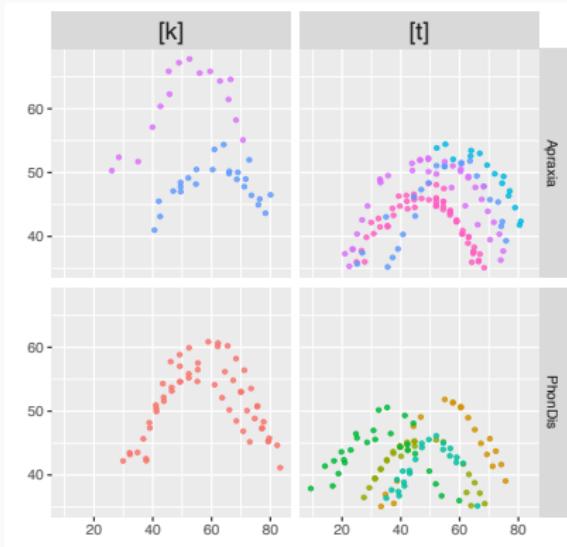


Figure 3: Dados individuais de ultrassom

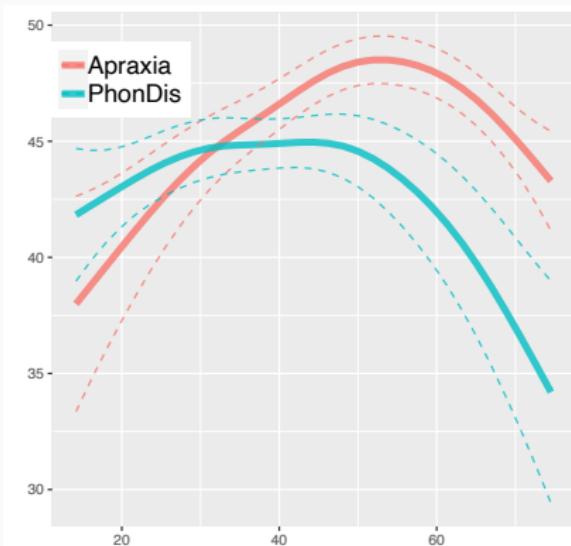


Figure 4: Smoothing Splines

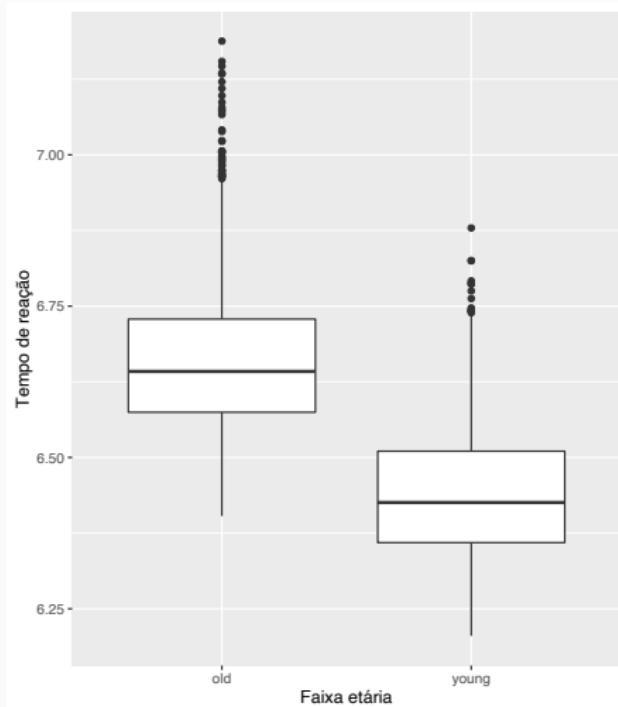
RStudio

Modelo com previsor categórico

Modelo com previsor categórico

Dados 'english' do pacote
LanguageR

- Criar um modelo para estimar **tempo de reação** em função da **faixa etária**



Modelo com previsor categórico

Criar um modelo para estimar **tempo de reação** ~ **faixa etária**

```
1 modEnAge = lm(RTlexdec ~ AgeSubject, data = en)
2 summary(modEnAge)

3 Coefficients:
4             Estimate Std. Error t value Pr(>|t|)
5 (Intercept)  6.660958  0.002324 2866.44   <2e-16 ***
6 AgeSubjectyoung -0.221721  0.003286  -67.47   <2e-16 ***
```

→ O output apresenta um *intercept* e uma *slope*, mas não temos uma reta cruzando o eixo y para termos um ponto de interseção e um ângulo

Modelo com previsor categórico

Criar um modelo para estimar **tempo de reação** ~ **faixa etária**

```
1 modEnAge = lm(RTlexdec ~ AgeSubject, data = en)
2 summary(modEnAge)
3 
4 Coefficients:
5             Estimate Std. Error t value Pr(>|t|)
6 (Intercept) 6.660958  0.002324 2866.44 <2e-16 ***
7 AgeSubjectyoung -0.221721  0.003286 -67.47 <2e-16 ***
8
```

- **intercept** = valor do tempo de reação quando `AgeSubject` está no valor de referência, que é o primeiro na ordem alfabética, neste caso 'old'
- **slope** = quanto muda no tempo de reação ao mudar o `AgeSubject` para 'young'

$$\beta_0 = 6.66 \text{ valor médio de TR para 'old'}$$

$$\beta_1 = 0.22 \implies 6.66 - 0.22 = 6.44 \text{ valor médio de TR para 'young'}$$

Modelo com previsor categórico

6.66 valor médio de TR para 'old'

6.44 valor médio de TR para 'young'

→ São exatamente os valores obtidos para as médias dos grupos com um teste-t

```
1 t.test(RTlexdec ~ AgeSubject, data = en)
2 
3 data: RTlexdec by AgeSubject
4 t = 67.468, df = 4534.6, p-value < 2.2e-16
5 alternative hypothesis: true difference in means is not equal to 0
6 95 percent confidence interval:
7   0.2152787 0.2281642
8 sample estimates:
9   mean in group old mean in group young
          6.660958           6.439237
```

→ Por que criar um modelo em vez rodar um test-t?

Modelo com previsor categórico

Vantagens de usar um modelo de regressão em vez rodar um test-t (ou uma ANOVA para previsores categóricos com mais de 2 níveis)

- visualizar o resultado da estimativa em termos da diferença entre os níveis (tamanho do efeito)
 - Os testes-t e valores-p dos coeficientes testam a H₀ de que eles são zero
- o modelo linear apresenta outras informações importantes:
 - resíduos (quanto dos dados o modelo não é capaz de estimar)
 - R^2 para ver quanto de y o modelo é capaz de prever
 - R^2 e F-statistic para comparação de modelos
- **a mais importante:** a possibilidade de se adicionar mais de uma variável previsora (várias delas!), entre elas numéricas e categóricas, e comparar modelos diferentes (como veremos em regressões múltiplas)

RStudio