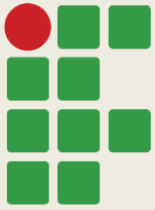
Probabilidade e Estatística



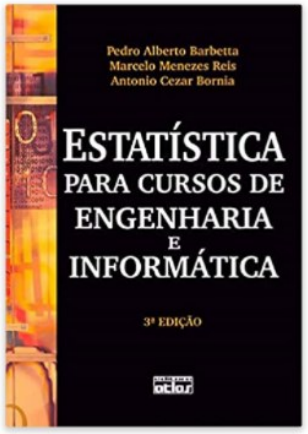
INSTITUTO FEDERAL
Catarinense
Campus Blumenau

Professor Jeovani Schmitt

Probabilidade e Estatística

Correlação e Regressão Linear Simples

- Correlação
- Coeficiente de correlação linear de Pearson
- Regressão linear simples



Capítulo 11, pp. 316 - 351

Introdução

XI SIMPROD
XI SIMPÓSIO DE ENGENHARIA DE PRODUÇÃO DE SERGIPE

TEMA:
"A ENGENHARIA DE PRODUÇÃO COMO MEIO DE TRANSFORMAÇÃO SOCIAL."
18 A 22 DE NOVEMBRO 2019

O efeito da publicidade em uma empresa: uma análise estatística

SANTOS, Pedro Vieira Souza*

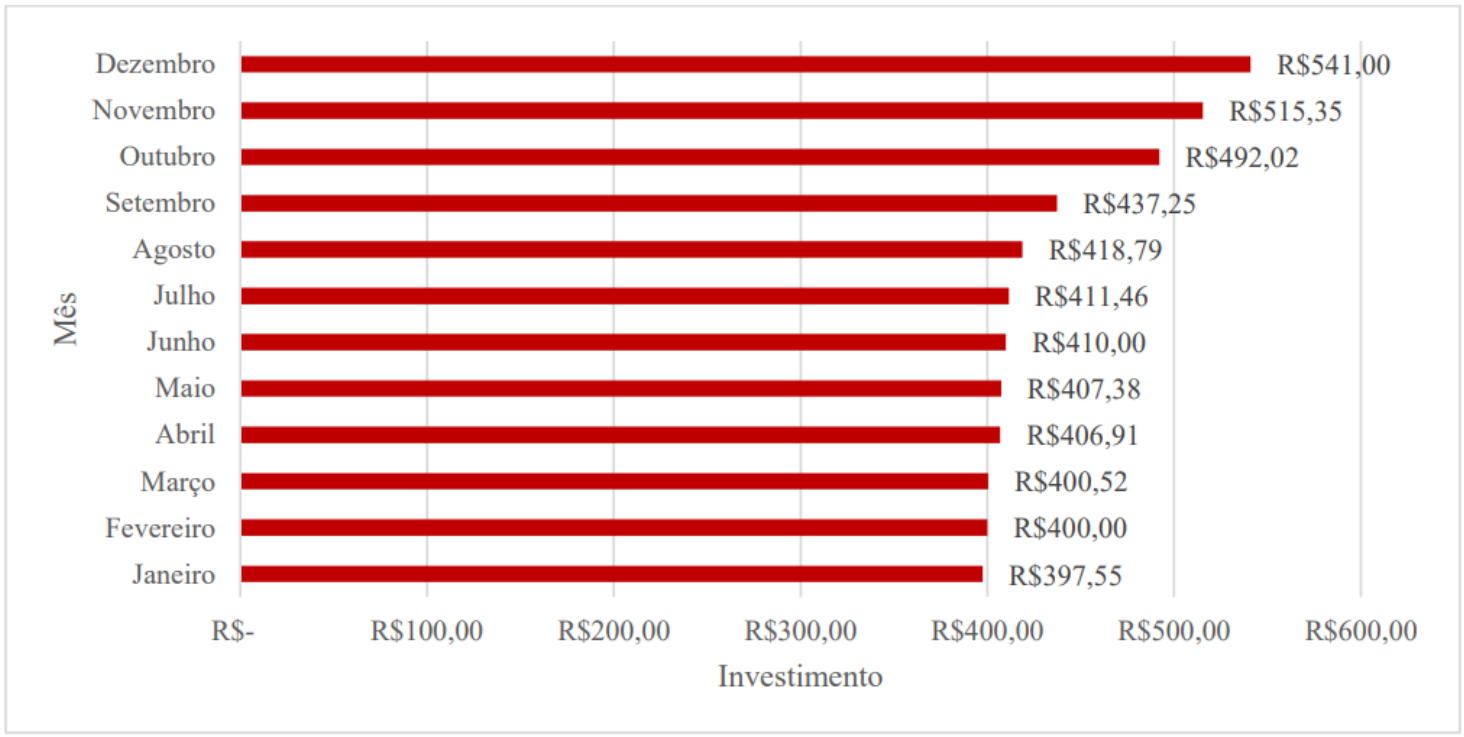
Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Pernambuco, Campus Caruaru – UFPE;
* Autor de correspondência. E-mail: pedrovieirass@hotmail.com

SANTOS, Pedro Vieira Souza. O efeito da publicidade em uma empresa: uma análise estatística. *In:* XI SIMPÓSIO DE ENGENHARIA DE SERGIPE, 2019. **Anais** do XI Simpósio de Engenharia de Produção de Sergipe. Disponível em: https://ri.ufs.br/bitstream/riufs/12579/2/EfeitoPublicidadeAnaliseEstatistica.pdf. Acesso em m19 maio 2023.

Investimento em propaganda x vendas

O quanto o investimento em propagandas aumenta as vendas ao longo do mês?

Figura 1 – Investimento em propaganda ao longo de 2017



Fonte: Dados da empresa (2018)

SANTOS, Pedro Vieira Souza. O efeito da publicidade em uma empresa: uma análise estatística. *In:* XI SIMPÓSIO DE ENGENHARIA DE SERGIPE, 2019. **Anais** do XI Simpósio de Engenharia de Produção de Sergipe. Disponível em: https://ri.ufs.br/bitstream/riufs/12579/2/EfeitoPublicidadeAnaliseEstatistica.pdf. Acesso em m19 maio 2023.

Investimento em propaganda x vendas

Tabela 3 – Investimento em propaganda *versus* vendas

Mês	Investimento em propaganda		Vendas
Janeiro	R$	397,55	R$17.607,03
Fevereiro	R$	400,00	R$18.699,45
Março	R$	400,52	R$17.991,30
Abril	R$	406,91	R$18.209,77
Maio	R$	407,38	R$18.734,00
Junho	R$	410,00	R$23.179,06
Julho	R$	411,46	R$21.334,80
Agosto	R$	418,79	R$21.847,13
Setembro	R$	437,25	R$21.009,48
Outubro	R$	492,02	R$21.157,09
Novembro	R$	515,35	R$22.654,33
Dezembro	R$	541,00	R$23.008,96

Fonte: Dados da empresa (2018)

arquivo: propaganda.xlxs

Correlação

Objetivo: Estamos estudando um problema em que temos interesse em analisar o comportamento conjunto de duas variáveis quantitativas.

Exemplos:

- Tempo de prática de esporte e ritmo cardíaco
- Tempo de estudo e nota na prova
- Taxa de desemprego e taxa de criminalidade
- Expectativa de vida e taxa de analfabetismo
- Vendas e Gasto com publicidade
- Número de clientes e vendas de uma empresa.

Correlação - Introdução

- Interesse em obter uma **medida estatística** que indique se existe ou não uma associação linear entre duas variáveis; e se existe, qual a sua magnitude e sinal.

  - Exemplos:

  1) anos de escolaridade e renda;

  2) medida da pressão arterial e idade.

Pressupostos básicos

- Os dados provêm de observações emparelhadas:
  - Peso e altura das ***mesmas*** crianças, medidas na ***mesma*** época (pares de medidas para cada criança).
  - Número de clientes e vendas de uma empresa no ***mesmo*** mês (pares de medidas para cada mês).

Pressupostos básicos

- Variáveis QUANTITATIVAS (ou assumidas quantitativas).
- Há apenas UMA variável dependente (de resposta).
- Supõe-se que os dados são oriundos de uma amostra aleatória
- Tamanho da amostra utilizada deve ser razoável ($n > 30$) para garantir a confiabilidade das conclusões obtidas.

Correlação - introdução

- Análise bidimensional (diagrama de dispersão)

- Coeficiente de correlação linear de Pearson

  Resume o relacionamento entre duas variáveis quantitativas em apenas um **número**

- Modelo de regressão linear simples

  Descreve o relacionamento entre duas variáveis por meio de uma equação

Correlação e Regressão

- Análise de Correlação fornece um número que resume o relacionamento entre as variáveis, indicando a ***força*** e a ***direção*** do relacionamento.

- Análise de Regressão fornece uma equação matemática que descreve a ***natureza*** do relacionamento entre as duas variáveis, permitindo inclusive que sejam feitas previsões dos valores de uma delas em função dos valores das outras.

Diagrama de Dispersão

Objetivos

Fornecer uma ideia inicial de como duas variáveis quantitativas estão relacionadas

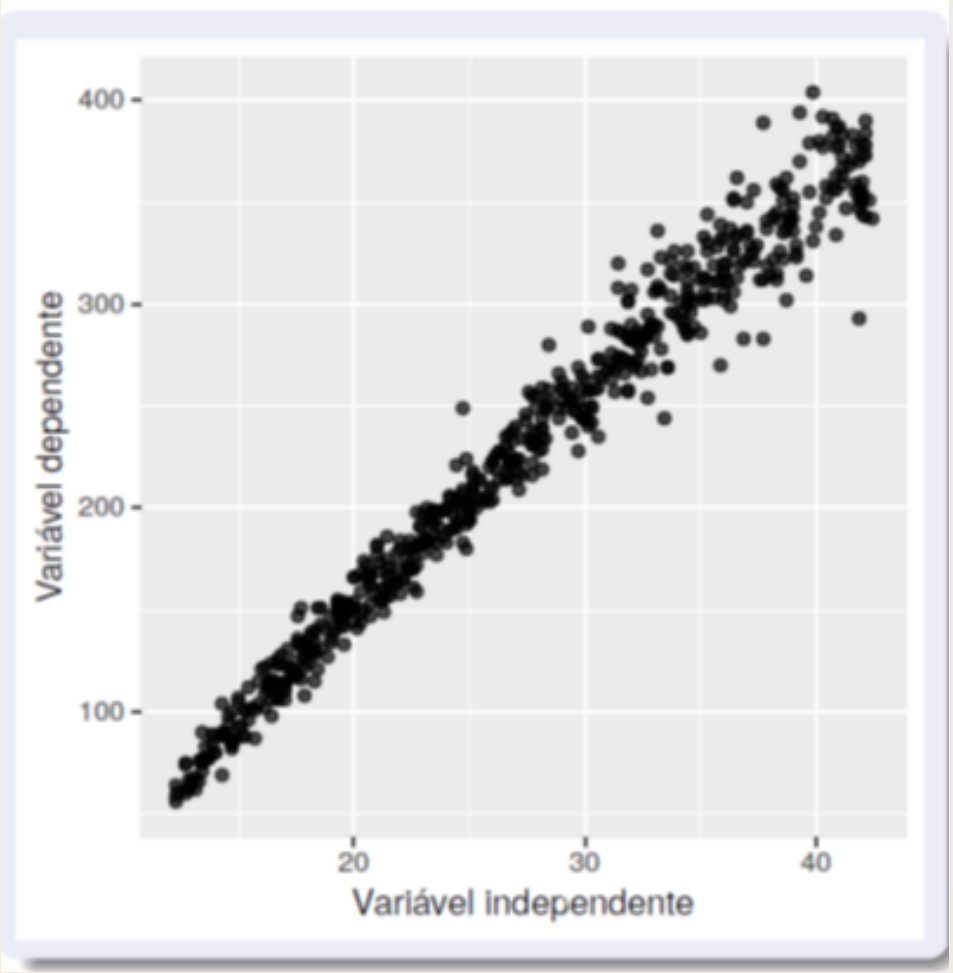
A **direção** dessa relação: o que acontece com Y quando X aumenta?

A **força** dessa relação: a qual "taxa" os valores de Y aumentam ou diminuem em função de X?

A **natureza** dessa relação: qual o tipo de relacionamento entre as duas variáveis? Podemos descrevê-lo com um reta, parábola, exponencial etc.

Diagrama de Dispersão

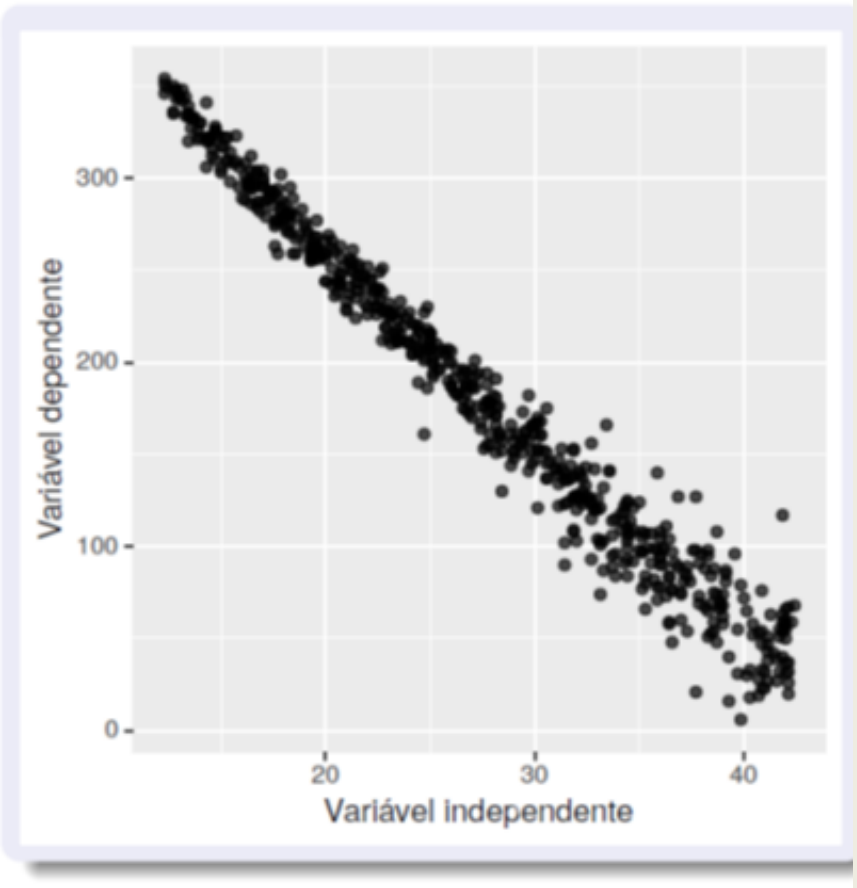
Exemplo 1



- **Direção:** à medida que a variável X aumenta, os valores de Y tendem a aumentar também
- **Força:** a taxa de crescimento é constante ao longo de todo eixo X
- **Natureza:** seria possível ajustar uma reta crescente que passasse por entre os pontos
- **Conclusão:** há correlação linear forte e positiva

Diagrama de Dispersão

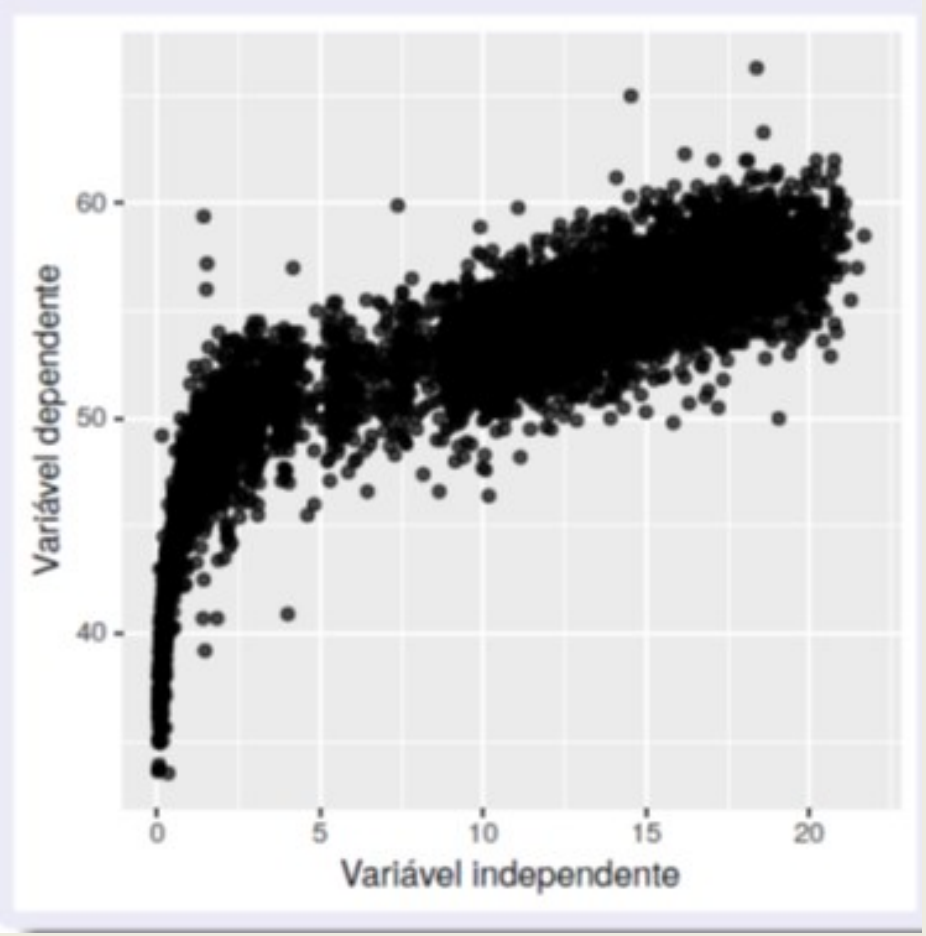
Exemplo 2



- **Direção:** à medida que a variável X aumenta, os valores de Y tendem a diminuir
- **Força:** a taxa de decrescimento é constante ao longo de todo eixo X
- **Natureza:** seria possível ajustar uma reta decrescente que passasse por entre os pontos
- **Conclusão:** há correlação linear forte e negativa

Diagrama de Dispersão

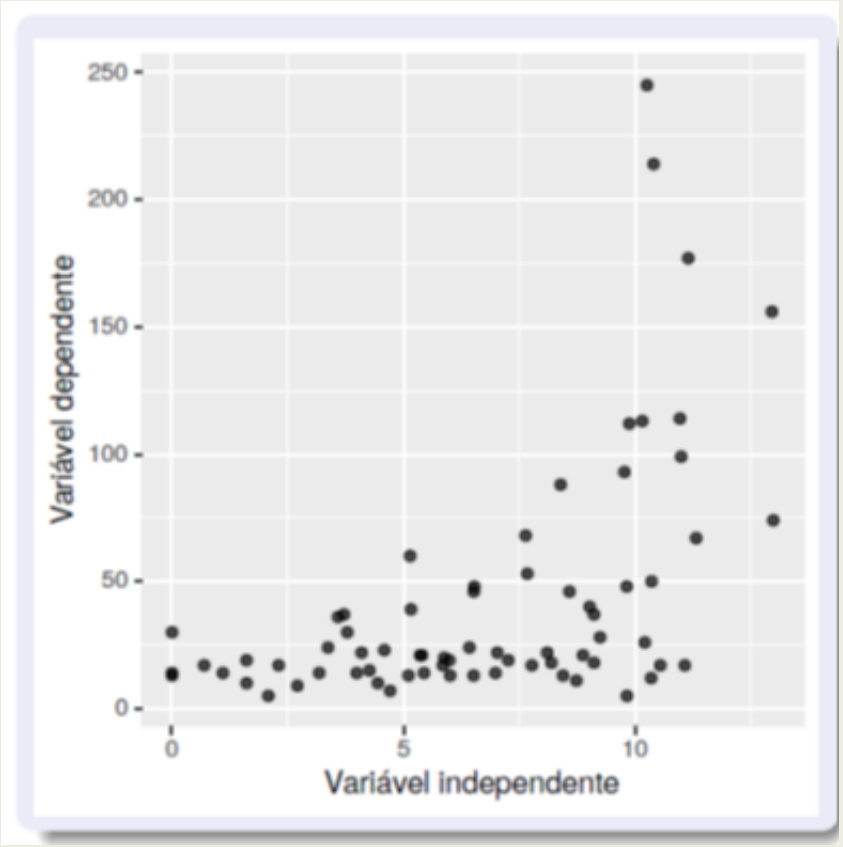
Exemplo 3



- **Direção:** à medida que a variável X aumenta, os valores de Y tendem a aumentar também
- **Força:** para valores muito pequenos de X a taxa de aumento em Y é muito alta. Posteriormente essa taxa é bem baixa, isto é, Y cresce de maneira extremamente suave.
- **Natureza:** não seria razoável ajustar uma reta que passasse por entre os pontos. Uma opção seria, por exemplo, utilizar uma função logarítmica.
- **Conclusão:** há forte correlação entre as variáveis, porém ela não é linear.

Diagrama de Dispersão

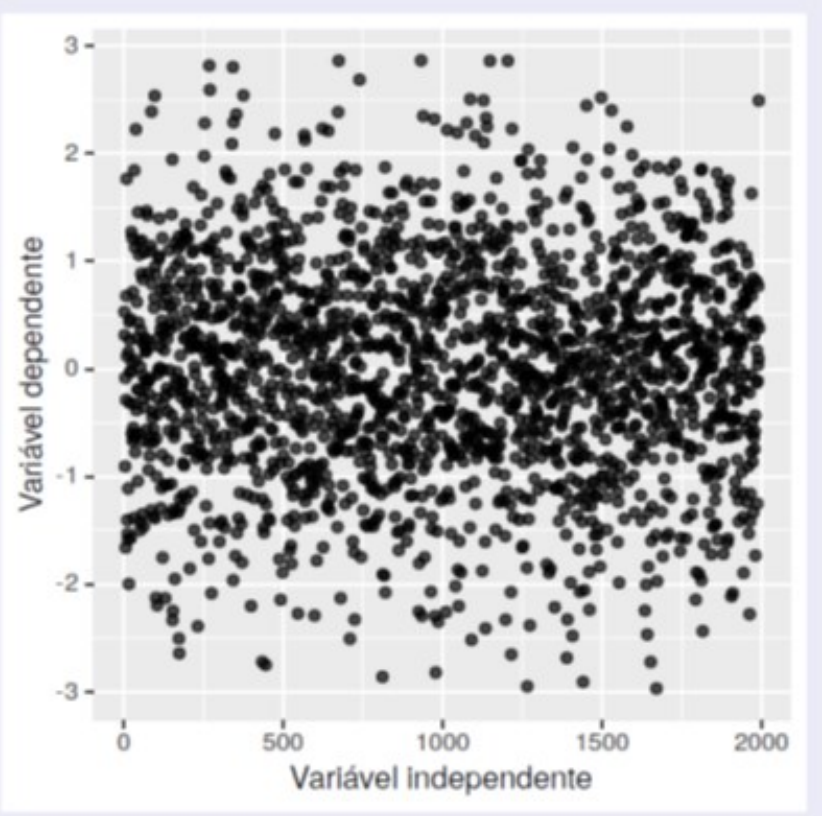
Exemplo 4



- **Direção:** à medida que a variável X aumenta, os valores de Y tendem a aumentar também
- **Força:** para valores pequenos de X a taxa de aumento em Y é quase nula. Posteriormente essa taxa aumenta, isto é, Y cresce de maneira mais acentuada.
- **Natureza:** não seria razoável ajustar uma reta que passasse por entre os pontos. Uma opção seria, por exemplo, utilizar uma função exponencial.
- **Conclusão:** há uma baixa ou moderada correlação entre as variáveis e ela não é linear

Diagrama de Dispersão

Exemplo 5



- **Direção:** não há um padrão aparente nos pontos
- **Força:** os pontos parecem se distribuir de maneira aleatória
- **Natureza:** não é possível considerar qualquer função para representar as observações
- **Conclusão:** não há um relacionamento aparente entre as duas variáveis

Coeficiente de correlação

Objetivo

Os objetivos do coeficiente de correlação linear de Pearson são o de mensurar, por meio de um único valor, o grau de relacionamento entre duas variáveis quantitativas, bem como indicar a direção dessa relação.

Notação

- O coeficiente de correlação linear de Pearson **populacional** é definido pela letra ρ
- O coeficiente de correlação linear de Pearson **amostral** é definido pela letra r

Correlação linear

O coeficiente de correlação linear amostral de Pearson:

$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n(\bar{x})^2}\sqrt{\sum_{i=1}^{n} Y_i^2 - n(\bar{y})^2}}$$

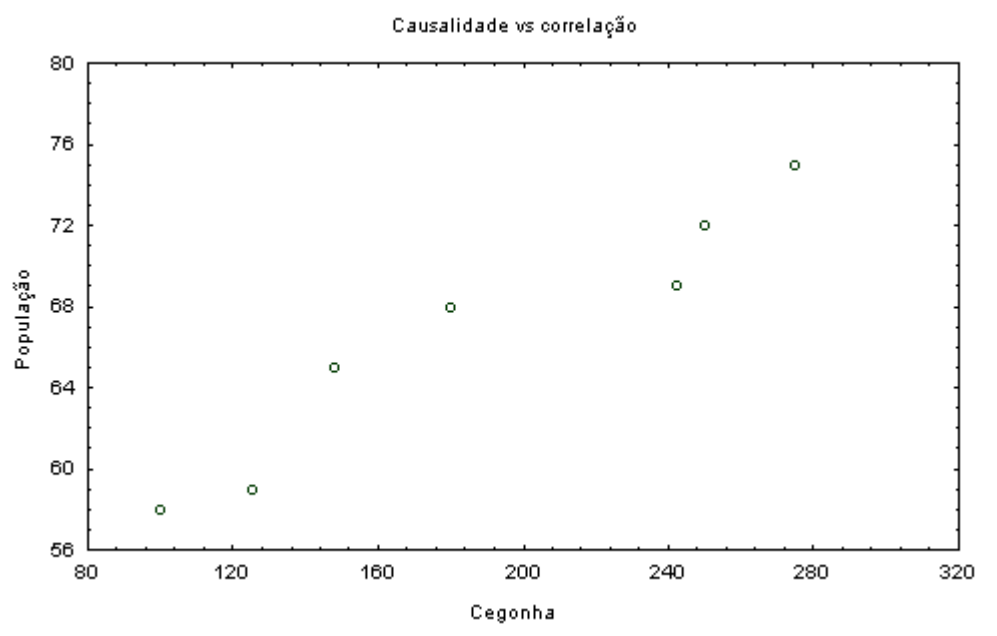
r – coeficiente de correlação linear amostral de Pearson;
n – número de pares de observações;
x – variável independente;
y – variável dependente.

Correlação linear - interpretação

Causalidade versos correlação

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* em ambas as situações não necessariamente implica numa relação de causa e efeito.

Exemplo: (Box, Hunter & Hunter, Statistics for Experimenters, p.8) O gráfico mostra a população de Oldemberg, Alemanha, no fim de cada um dos 7 anos (*Y)* contra o número de cegonhas (pássaros) naquele ano (*X).*



Interpretação: existe associação entre *X e Y.*

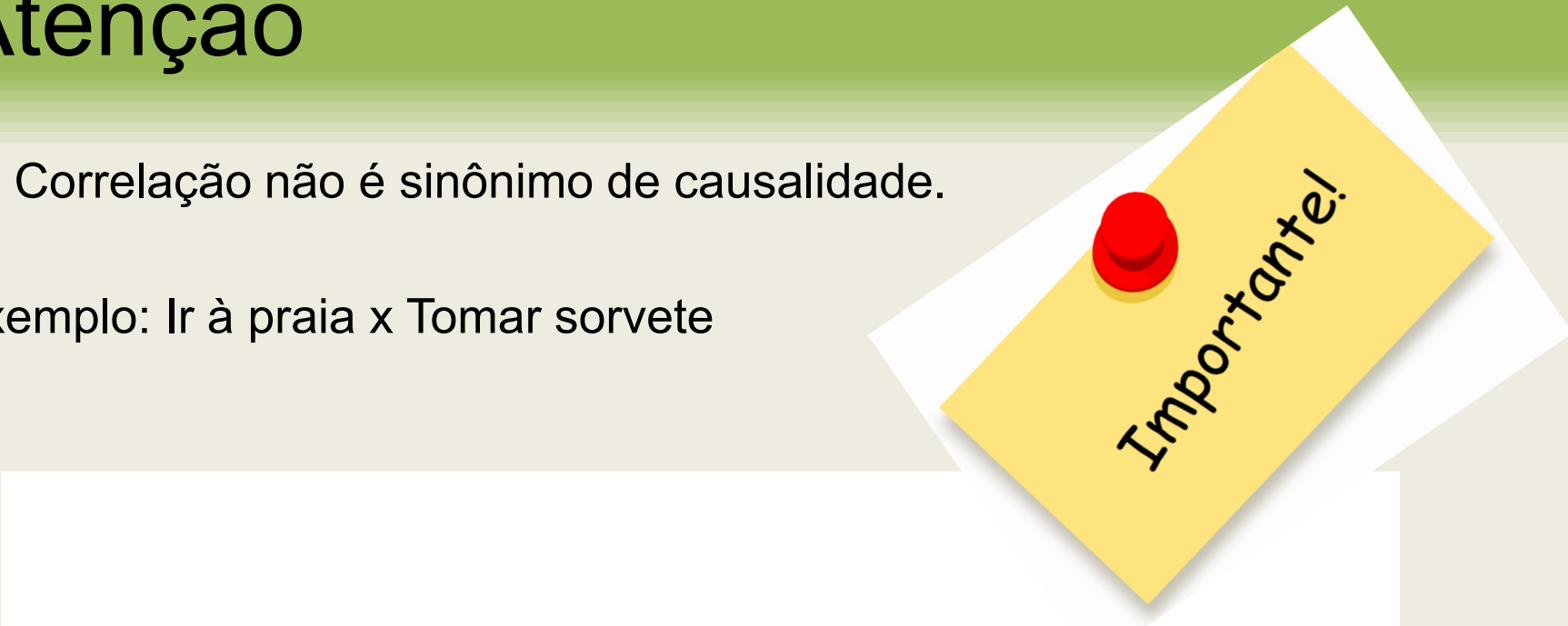
Frequentemente, quando duas *v. X e Y* parecem estar fortemente associadas, pode ser porque *X e Y* estão, de fato, associadas com uma terceira variável*, W.* No exemplo, *X e Y* aumentam com *W* = tempo.

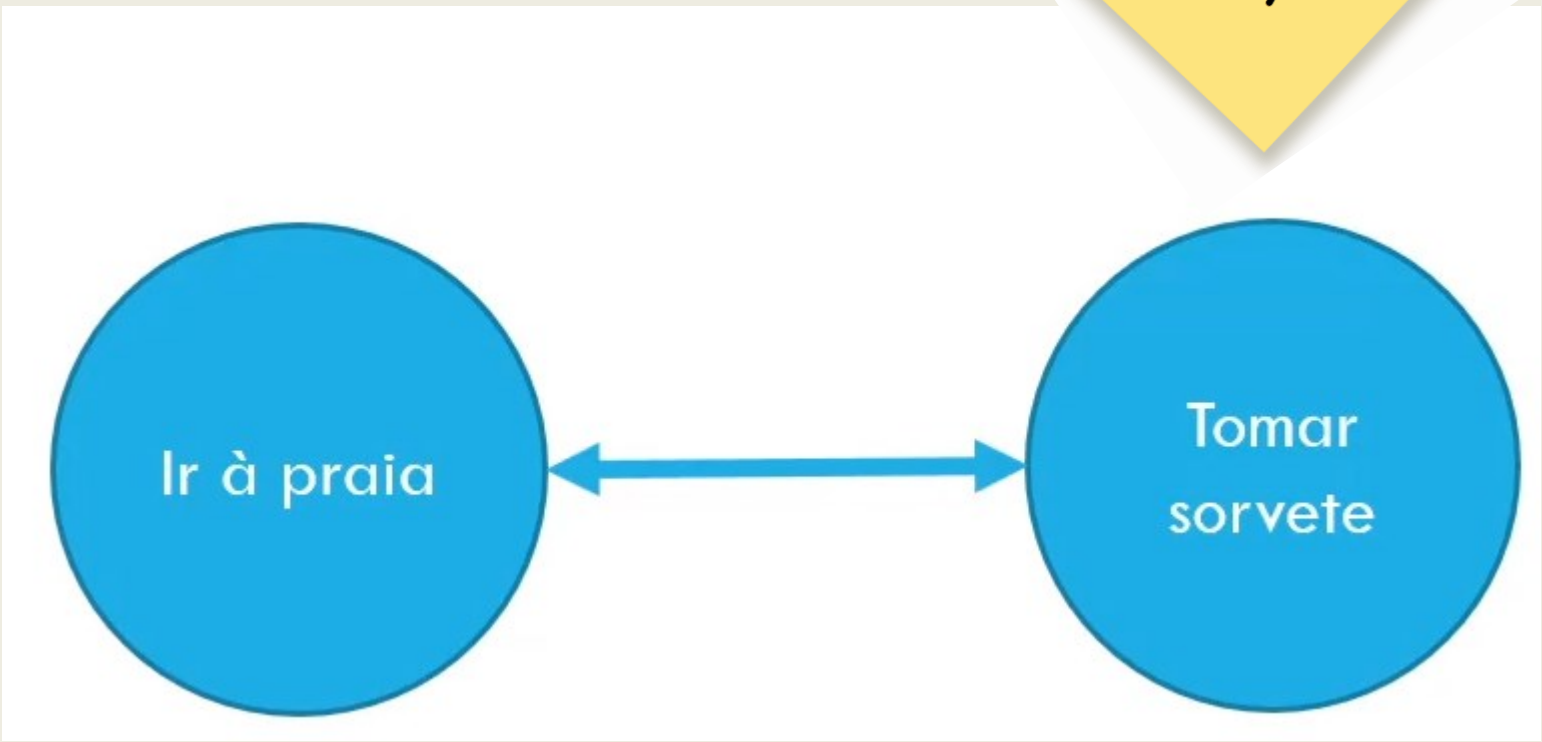
Correlação não necessariamente implica em causalidade

Atenção

- ✓ Correlação não é sinônimo de causalidade.

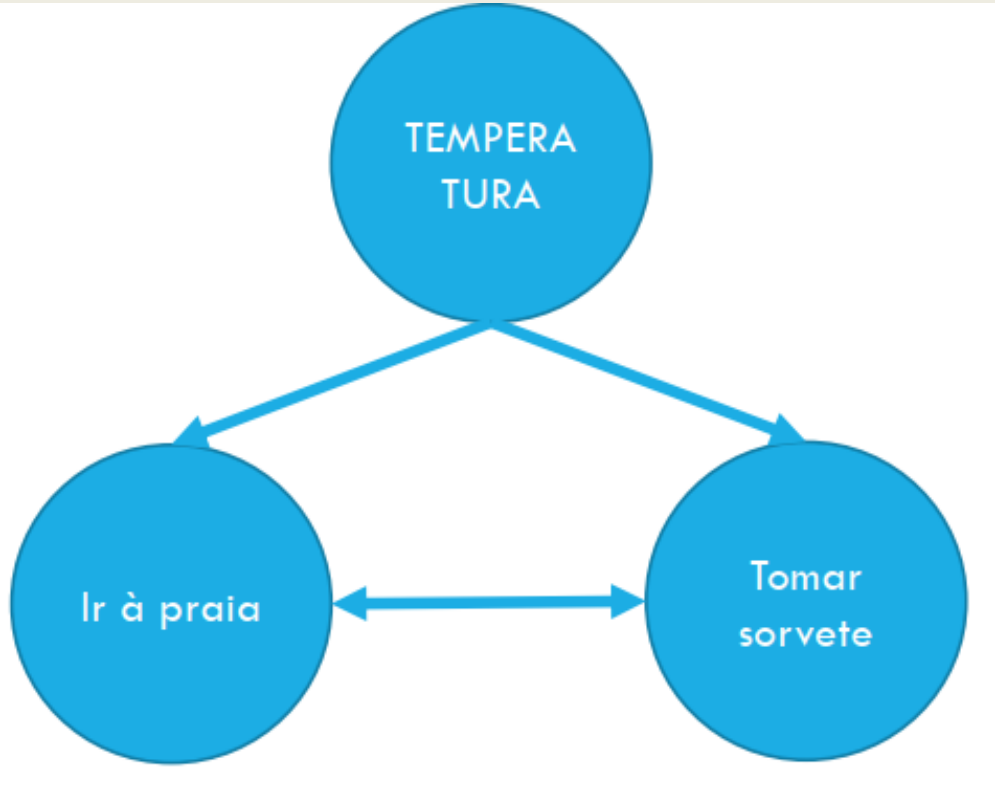
Exemplo: Ir à praia x Tomar sorvete





Atenção

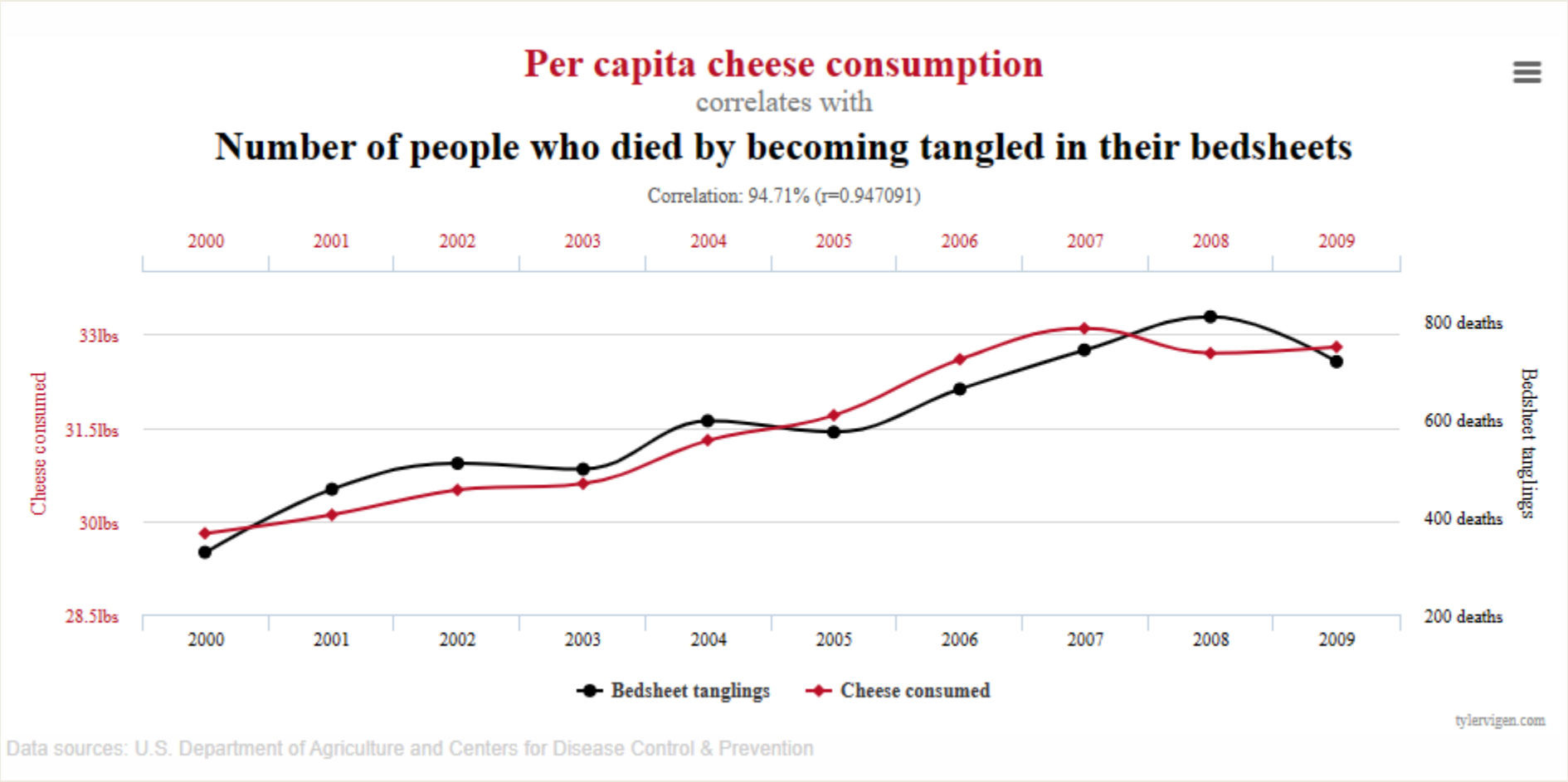
- ✓ A correlação entre duas variáveis pode ser causada por uma terceira variável oculta.



** É possível encontrar uma correlação completamente espúria entre duas variáveis.

Correlações espúrias

Link para visualizar correlações espúrias: https://tylervigen.com/spurious-correlations



Correlação linear – Exemplo 1

Os dados ao lado referem-se a idade e pressão arterial de 12 poessas. Construa um diagrama de dispersão e calcule o coeficiente de correlação linear de Pearson interpretando os resultados.

Idade	Pressão
56	147
42	125
72	160
36	118
47	128
55	150
49	145
38	115
42	140
68	152
60	155
63	149

Correlação linear – Exemplo 1

Sintaxe no R

```
# Dados
idade <- c(56, 42, 72, 36, 47, 55, 49, 38, 42, 68, 60, 63)
pressao <- c(147, 125, 160, 118, 128, 150, 145, 115, 140, 152, 155, 149)

#Gráfico
plot(idade, pressao)

# Gráfico pelo GGPLOT
library(ggplot2)
dados <- data.frame(idade, pressao)
ggplot(data = dados) +
geom_point(aes(x = idade, y = pressao))

# Coeficiente de correlação
cor(idade, pressao)
```

Correlação linear – Exemplo 2

- Processo de queima de massa cerâmica para pavimento
  - X_1 = retração linear (%),
  - X_2 = resistência mecânica (MPa) e
  - X_3 = absorção de água (%).

Pede-se: *Verifique se existe correlação linear entre as variáveis*

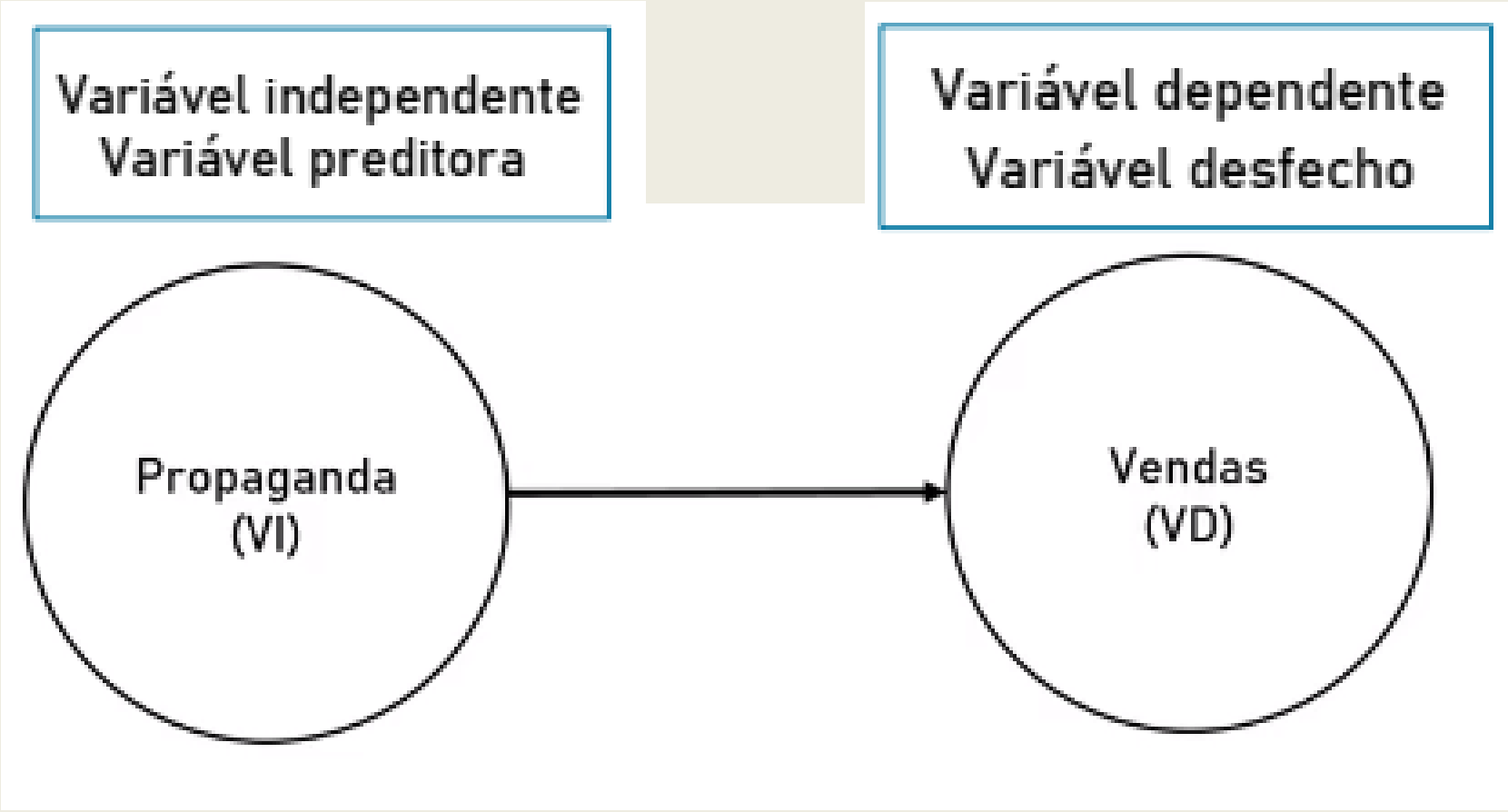
Faça os diagramas de dispersão para a relação entre as variáveis e calcule os coeficientes de correlação para os dados que estão no arquivo *ceramica.RData*

Análise de regressão

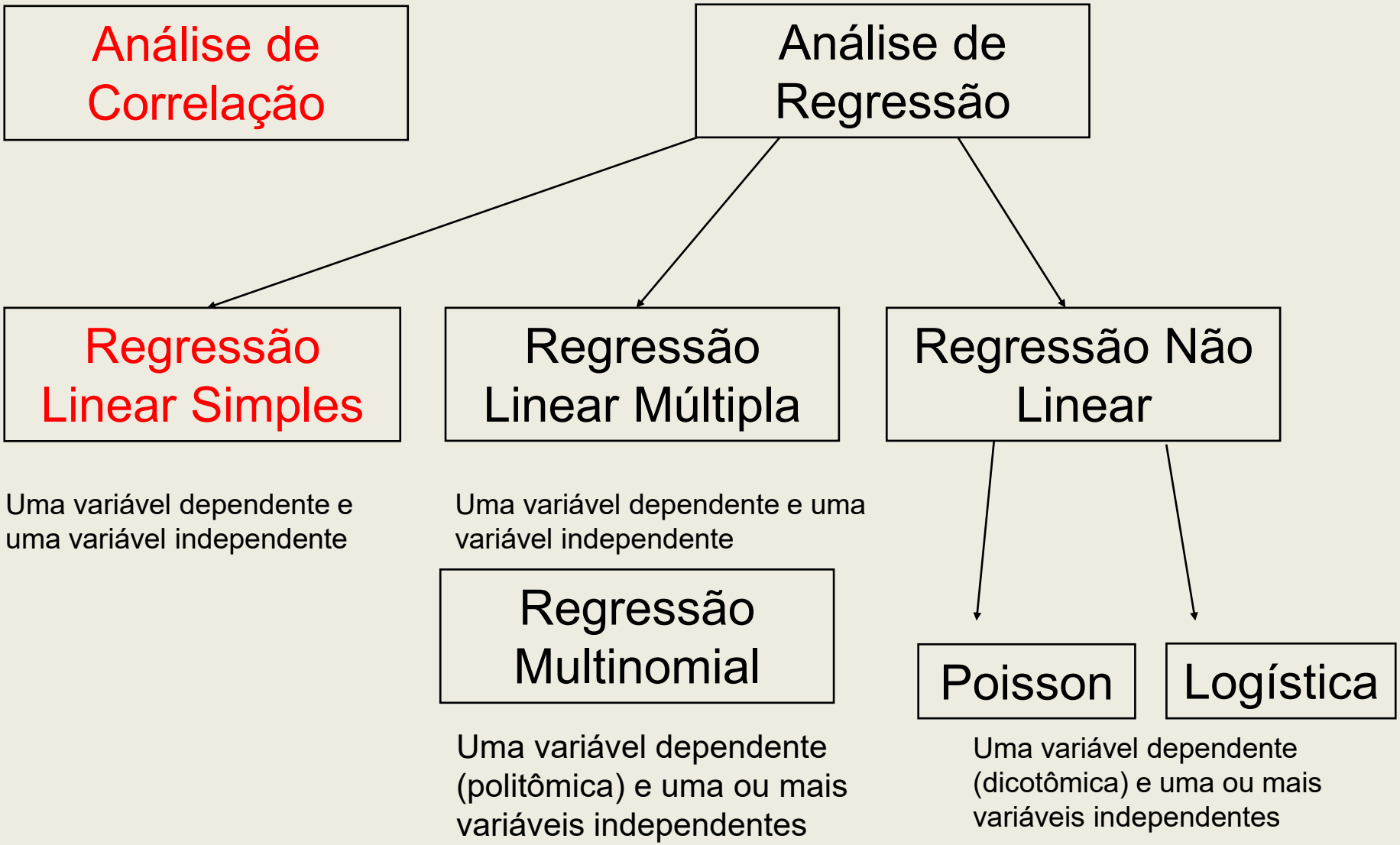
Análise de regressão é uma metodologia estatística que utiliza a ***relação*** entre duas ou mais variáveis quantitativas (ou qualitativas) de tal forma que uma variável pode ser predita a partir da outra ou outras.

Exemplo de análise de regressão

O quanto o investimento em propagandas aumenta as vendas ao longo do mês?

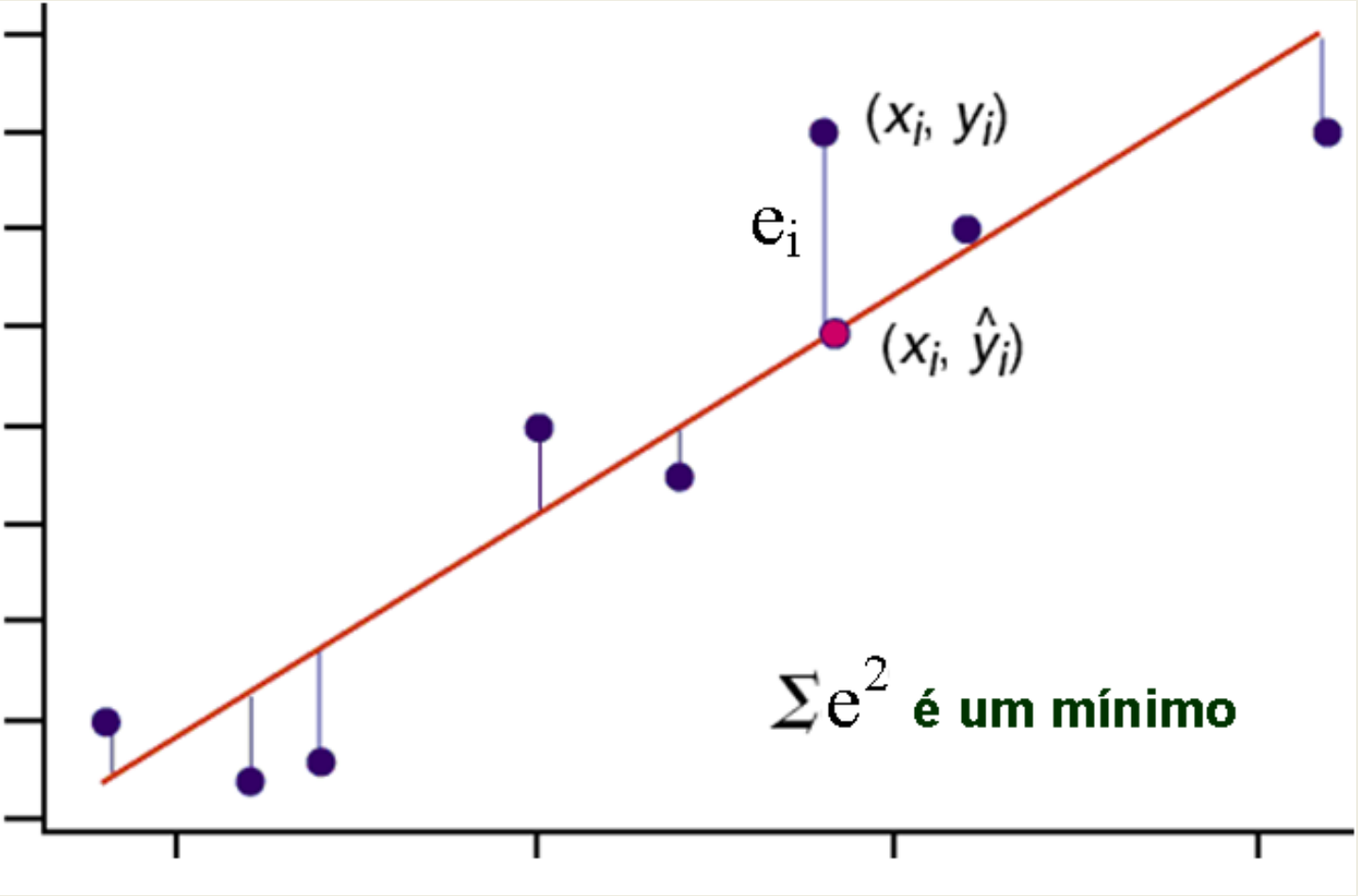


Classificação dos modelos



Regressão linear

A reta de regressão



Regressão linear

Modelo

$$Y = \left[\text{Predito por X, segundo uma função}\right] + \left[\text{Efeito aleatório}\right]$$

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Regressão Linear Simples

Parâmetros ou coeficientes de regressão

Com $i = 1,2,\ldots,n$

Regressão linear simples

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

Em que:

$Y = variável\ dependente$

$\beta_0 = intercepto\ (constante)$

$\beta_1 = o\ grau\ sobre\ o\ quanto\ X\ impacta\ Y$

$X = variável\ independente$

$e = erro\ (efeito\ aleatório)$

O modelo de regressão linear

- Estimativas dos parâmetros utilizando o método dos mínimos quadrados

Estimativa de β_1 :

$$b = \frac{n.\sum(x_i y_i) - (\sum x_i)\cdot(\sum y_i)}{n.\sum x_i^2 - (\sum x_i)^2}$$

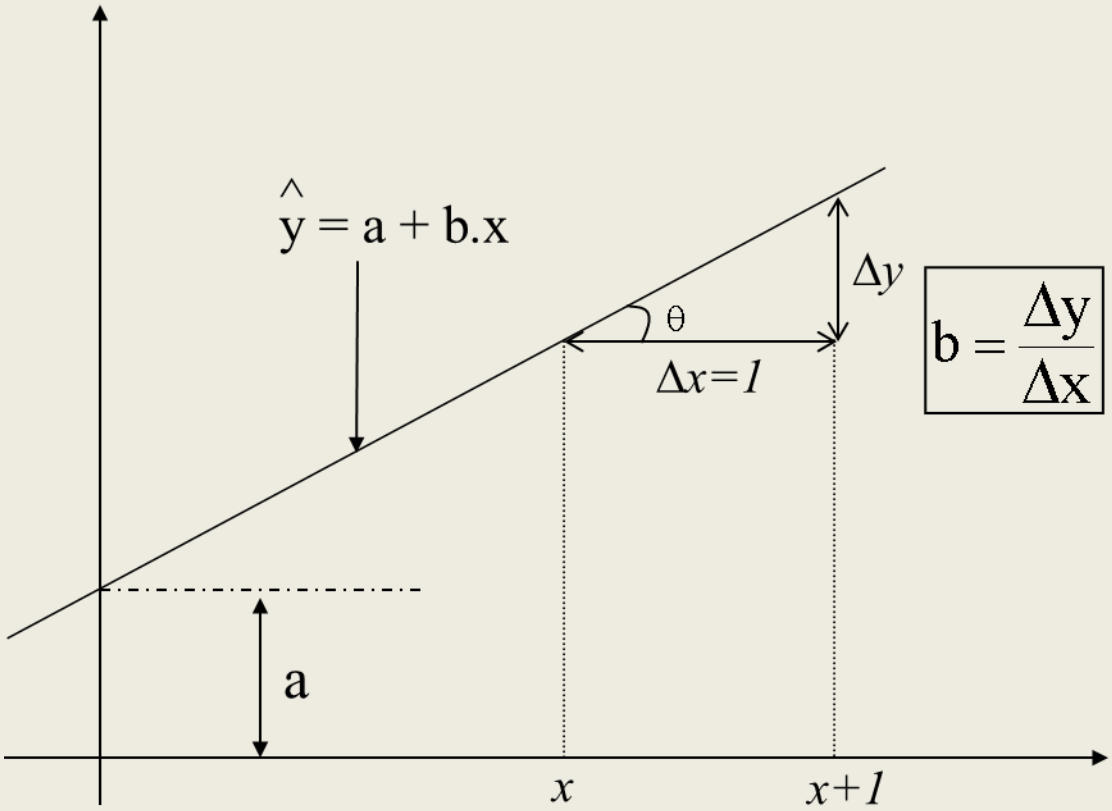
Estimativa de β_0 :

$$a = \frac{\sum y_i - b\sum x_i}{n}$$

Reta de regressão construída com os dados:

$$\hat{y}_i = a + bx_i$$

Interpretação dos parâmetros



a (intercepto) é o valor da média da distribuição de Y em X=0, não tem significado prático como um termo separado (isolado) no modelo;

b (inclinação) expressa a *taxa de mudança* em *Y,* isto é, é a mudança em *Y* quando ocorre a mudança de uma unidade em *X.*

Regressão linear simples – Exemplo 3

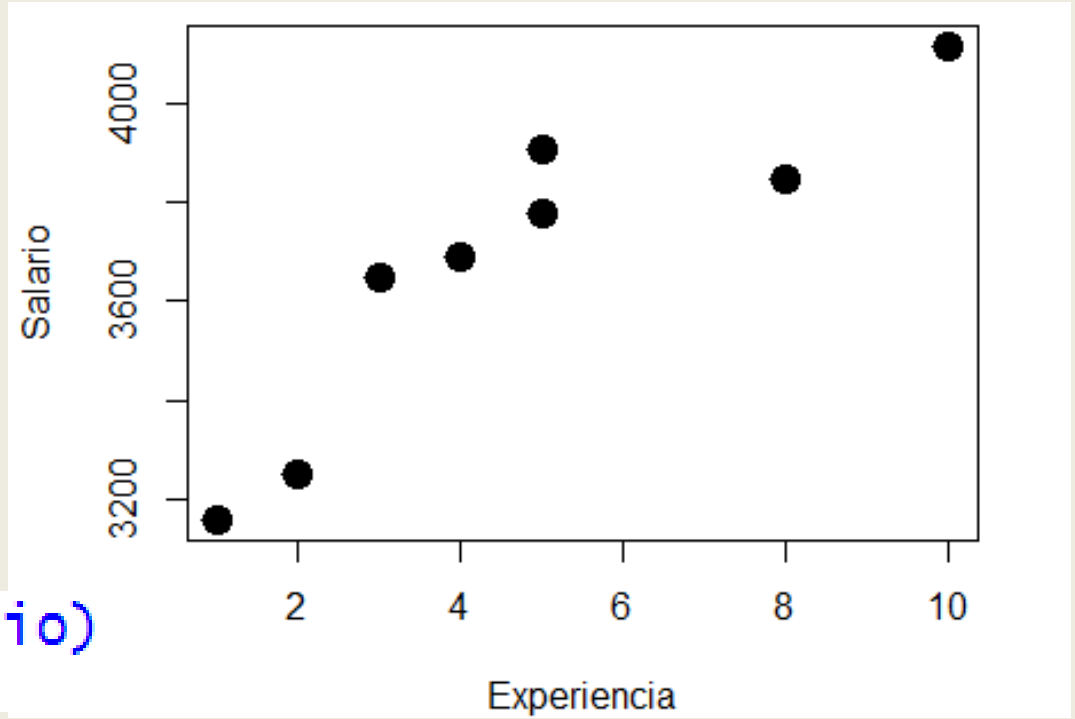
- Na tabela ao lado é dado o tempo de experiência e o salário de 8 colaboradores que exercem a mesma função em uma empresa.
- **X**: Experiência (anos de trabalho na empresa)
- **Y**: Salário (R$)

X	Y
1	3156
2	3248
3	3650
4	3689
5	3779
5	3907
8	3849
10	4118

Regressão linear simples – Exemplo 3

Sintaxe no R

```
Experiencia <- c(1,2,3,4,5,5,8,10)
Salario <- c(3156,3248,3650,3689,3779,3907,3849,4118)
cor(Experiencia, Salario)
plot(Experiencia, Salario, lwd=8)
```



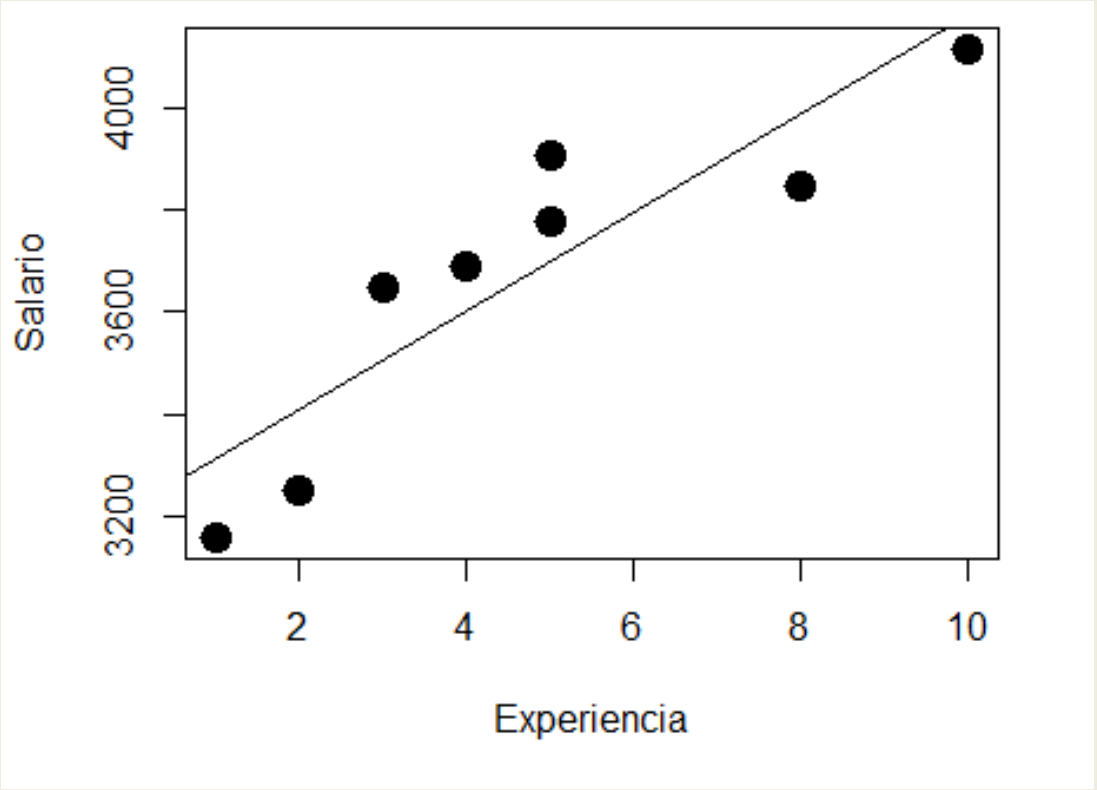

```
> cor(Experiencia, Salario)
[1] 0.8918903
```

Regressão linear simples – Exemplo 3

Sintaxe no R

```
# Coeficientes do modelo
lm(Salario ~ Experiencia)

# Reta de regressão no modelo
plot(Experiencia, Salario, lwd=8)
abline(lm(Salario ~ Experiencia))
```




```
Coefficients:
(Intercept)  Experiencia
     3216.03        96.52
```

Exemplo 3 – interpretação dos coeficientes

```
Coefficients:
(Intercept)  Experiencia
    3216.03        96.52
```

- O salário inicial de uma pessoa sem experiência é R$ 3.216,03

- A cada ano a mais de experiência, espera-se que o salário aumente R$ 96,52

$$\hat{y} = 3216{,}03 + 96{,}52.\, x$$

$$\widehat{salário} = 3216{,}03 + 96{,}52.\, Experiência$$

Regressão linear: Qualidade do ajuste

Coeficiente de determinação (R^2)

$$R^2 = \frac{\text{Variação explicada}}{\text{Variação total}} = \frac{\text{SQReg}}{\text{SQTotal}}$$

$$0 \leq R^2 \leq 1$$

Quanto mais alto é o valor de R^2, mais o modelo de regressão linear simples consegue explicar a variação de *Y*.

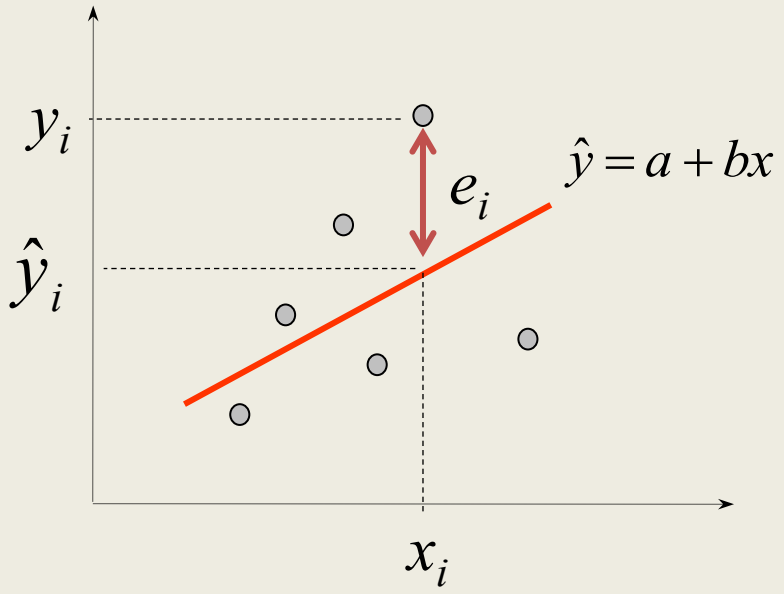
Análise de Resíduos

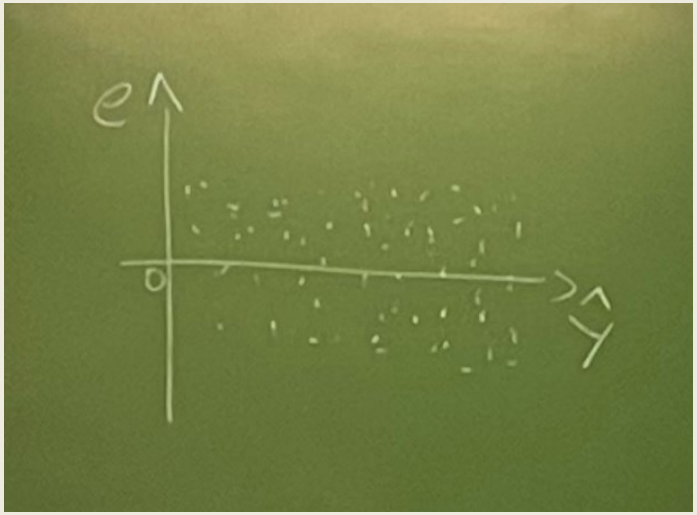
- Valores preditos:

$$\hat{y}_i = a + bx_i$$

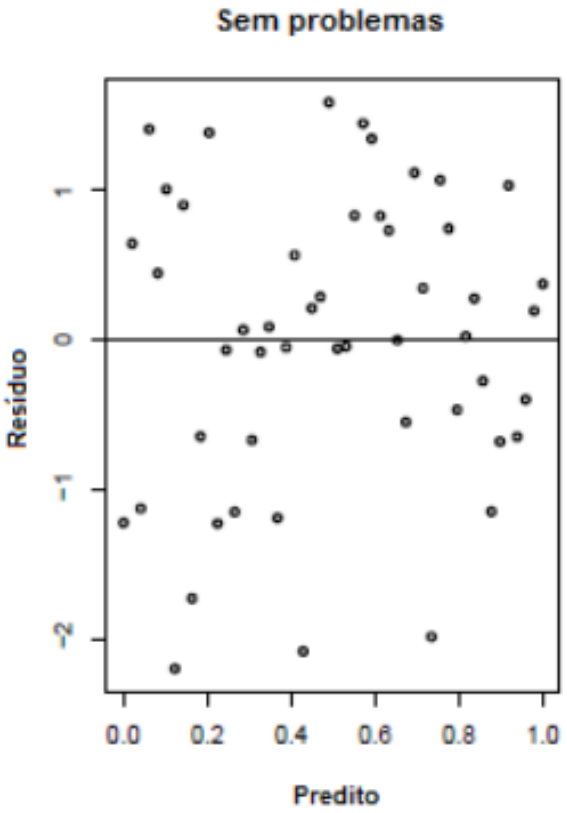
- Resíduos:

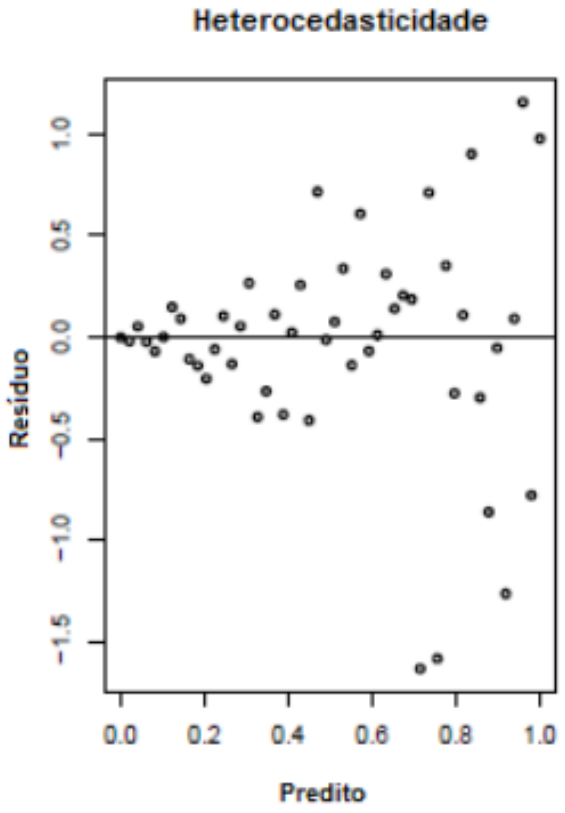
$$e_i = y_i - \hat{y}_i$$

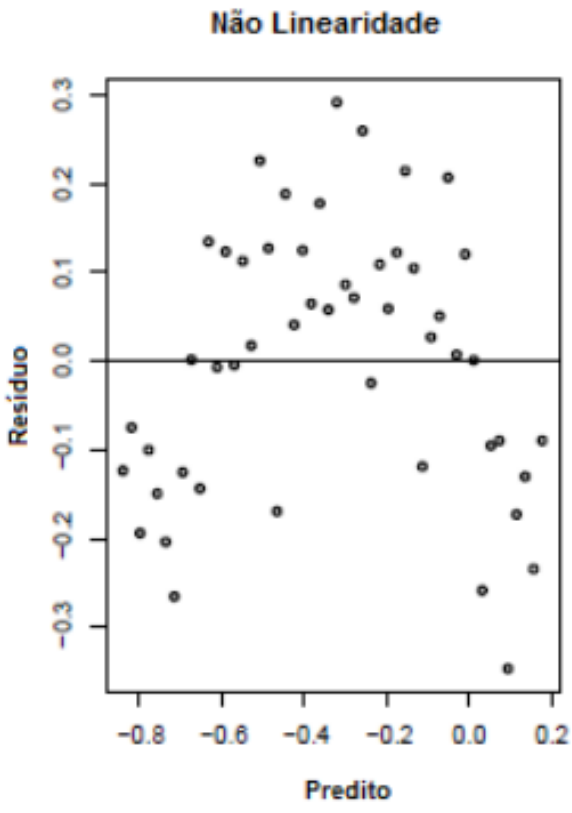




Análise de Resíduos







Análise de Resíduos

(a) Resíduos dispersos aleatoriamente em torno de zero, indica o comportamento esperado para distribuição dos erros

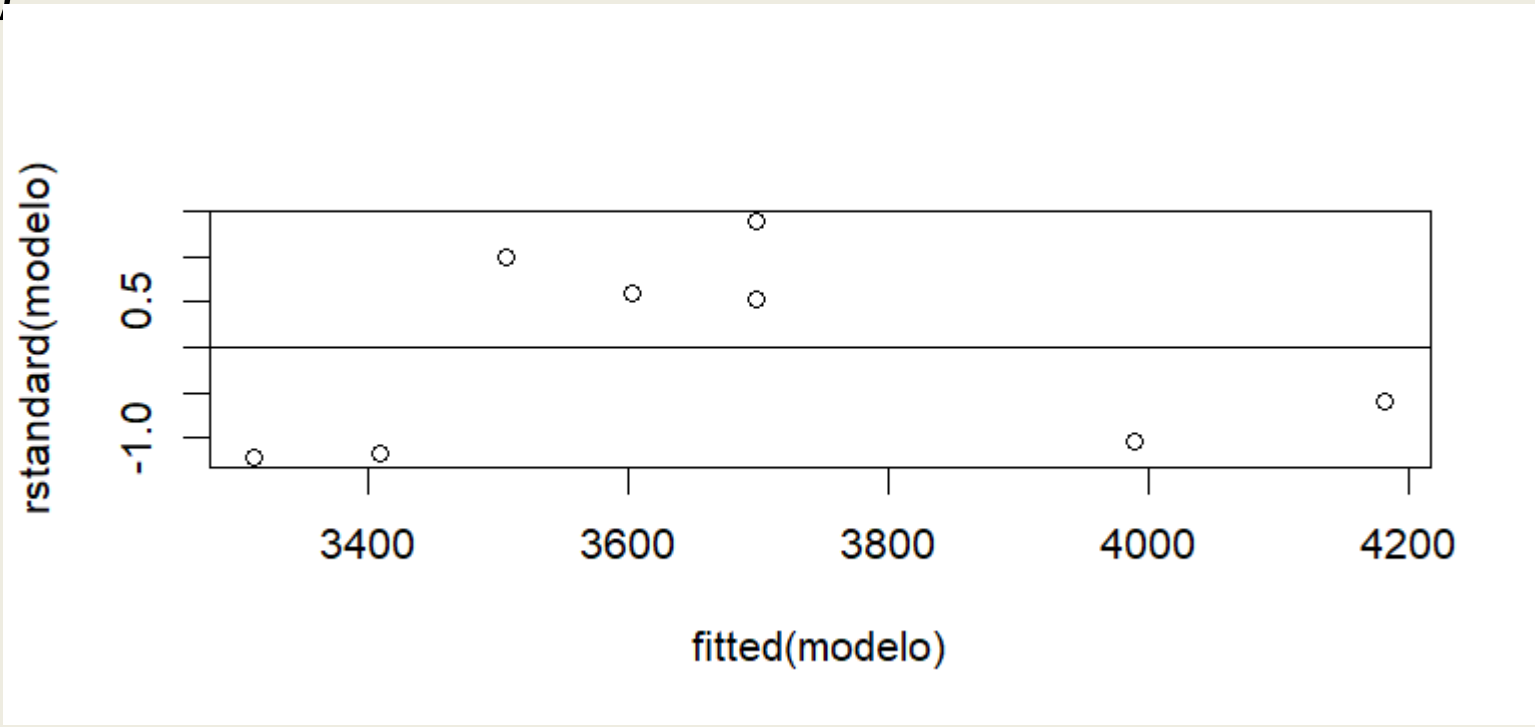
(b) Dispersão dos resíduos aumenta conforme o valor do predito, configurando heterogeneidade de variâncias dos erros (erros heterocedásticos); comum quando a variável resposta refere-se a contagens. Solução: transformar a variável resposta ou utilizar algum modelo linear generalizado

(c) Distribuição dos resíduos apresenta uma tendência não linear (no caso, quadrática). Solução: incorporar novas variáveis explicativas ao modelo, ou considerar alguma transformação em X e/ou Y, ou utilizar algum modelo de regressão não linear

Análise de Resíduos

Sintaxe no R

```
# Análise de resíduos
plot(fitted(modelo),
rstandard(modelo))
abline(0,0)
```



Coeficiente de determinação (R^2) – Exemplo 3

Sintaxe no R

```
# Coeficiente de determinação
summary(modelo)$r.squared
```

```
> summary(modelo)$r.squared
[1] 0.7954684
```

Interpretação do coeficiente de determinação R^2 = 79,5 %

79,5 % da variação dos salários é explicada pelo tempo de experiência.

20,5% da variação dos salários pode ser explicada por outras variáveis que influenciam no salário.

Estimação/Predição

- Estime o salário para uma pessoa com 7 anos de empresa

$$\hat{y} = 3216{,}03 + 96{,}52.\,x$$

$$\widehat{salário}(x = 7) = 3216{,}03 + 96{,}52.7$$

$$\widehat{salário}(x = 7) = R\$\ 3.891{,}67$$

Sintaxe no R

```
predict(modelo,data.frame(Experiencia=7))
```