

# MACHINE LEARNING



# Machine Learning

**Tema:** Regressão Linear

**Coordenação:**

Prof. Dr. Adolpho Walter  
Pimazzi Canton

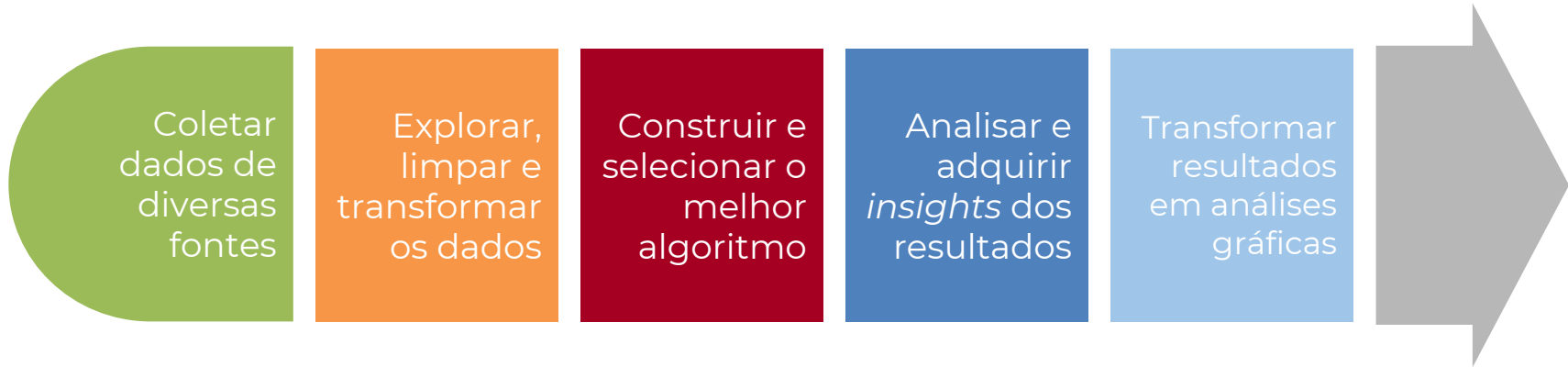
Profa. Dra. Alessandra de  
Ávila Montini

**Prof. Anderson França**

# Regressão

# Machine Learning e Regressão

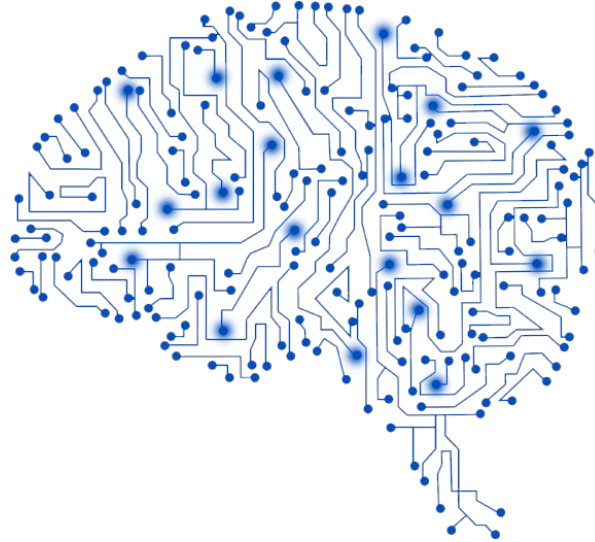
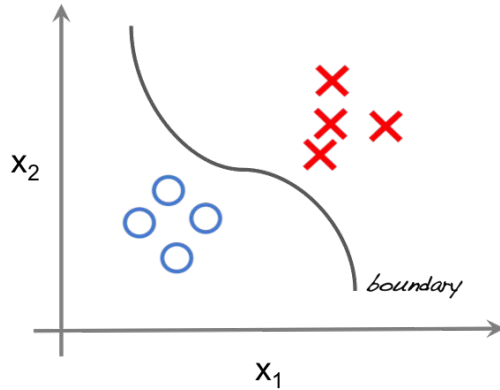
Aprendizado de Máquina (Machine Learning) é um campo de estudo que fornece a capacidade de uma Máquina de **entender dados** e **aprender com os dados**. O ML não é apenas sobre modelagem analítica, mas é uma modelagem de ponta a ponta que envolve as seguintes etapas:



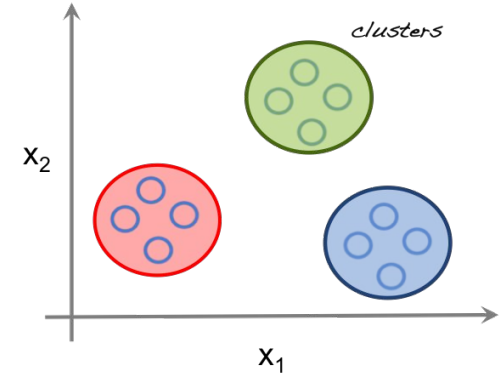
Fonte: [R-Bloggers](#)

# Machine Learning e Regressão

Supervised learning



Unsupervised learning



Modelos Supervisionados e

# Regressão

$$Y = f(x) + \varepsilon$$

**Y** = Define a variável Resposta

**f(x)** = define a função que depende do conjunto de recursos de entrada

**$\epsilon$**  = define o erro aleatório. Para o modelo ideal, deve ser aleatório e não deve depender de nenhuma entrada.

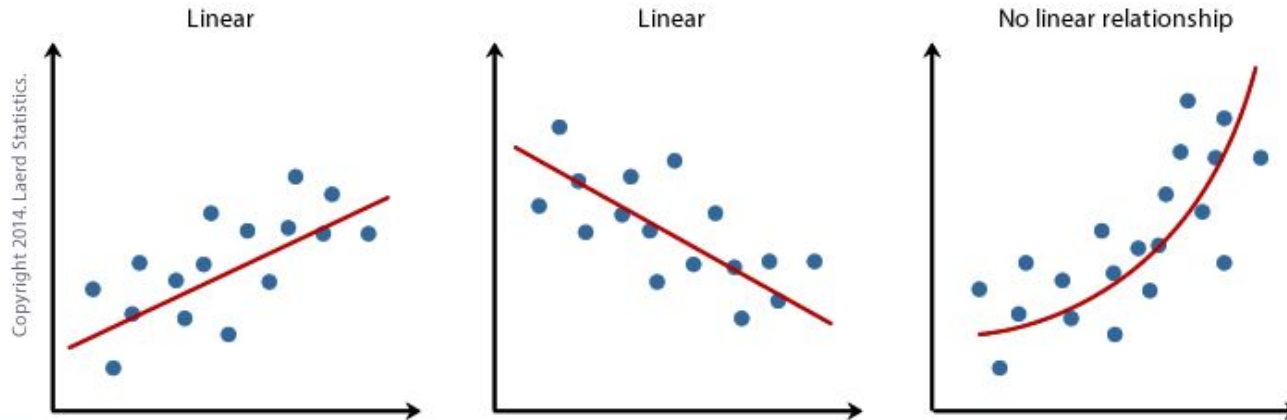
Equação básica para qualquer modelo Supervisionado

$$\mathbf{Y} = \mathbf{f(x)} + \epsilon$$

# Algoritmo Supervisionado e Regressão

A regressão linear é usada para prever o valor de uma variável de resultado  $Y$  (*outcome*) com base em uma ou mais variáveis preditoras de entrada  $X$ .

O objetivo é **estabelecer uma relação linear** (uma fórmula matemática) entre a(s) variável(is) preditora(s) e a variável resposta, dessa forma, podemos usar essa fórmula para estimar o valor da resposta  $Y$ , quando apenas os valores dos preditores ( $X$ s) são conhecidos.





# Algoritmo Supervisionado e Regressão

Na regressão linear, assumimos que a forma funcional,  **$f(\mathbf{x})$  é linear** e, portanto, podemos escrever a equação da seguinte forma:

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (\text{Regressão Simples})$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon \quad (\text{Regressão Múltipla})$$

Vamos simplificar

# Regressão Linear Simples

Tem o objetivo de projetar uma variável de interesse em função de uma variável auxiliar.

Por exemplo projetar o número de clientes que irão adquirir um cartão de crédito em função do número de benefícios oferecidos.



# Objetivo

Estudar a relação entre duas variáveis  
**quantitativas**

Projetar uma variável de interesse em função  
de uma variável auxiliar

# Aplicações

1. Projetar a venda de um produto em função do preço praticado
2. Projetar o salário anual em função do tempo de experiência em uma determinada empresa
3. Projetar a quantidade de produtos vendidos em função do investimento em mídia
4. Projetar a venda de ar-condicionado em função da temperatura

# Correlação Linear

Relação entre duas variáveis

# Coeficiente de Correlação

O coeficiente de correlação é uma medida descritiva da força da associação linear entre duas variáveis de escala métrica

Os valores do coeficiente de correlação estão sempre entre -1 e 1

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \cdot \text{var}(Y)}}$$

R

`cor(dados[,vars])`

Excel

**r** =CORREL(matrizY;matrizX)

# r : Coeficiente de correlação linear

$$-1 < r < 1$$

$r = 0$  **não existe correlação** linear entre as variáveis

$r = 1$  existe correlação linear **positiva perfeita** entre as variáveis

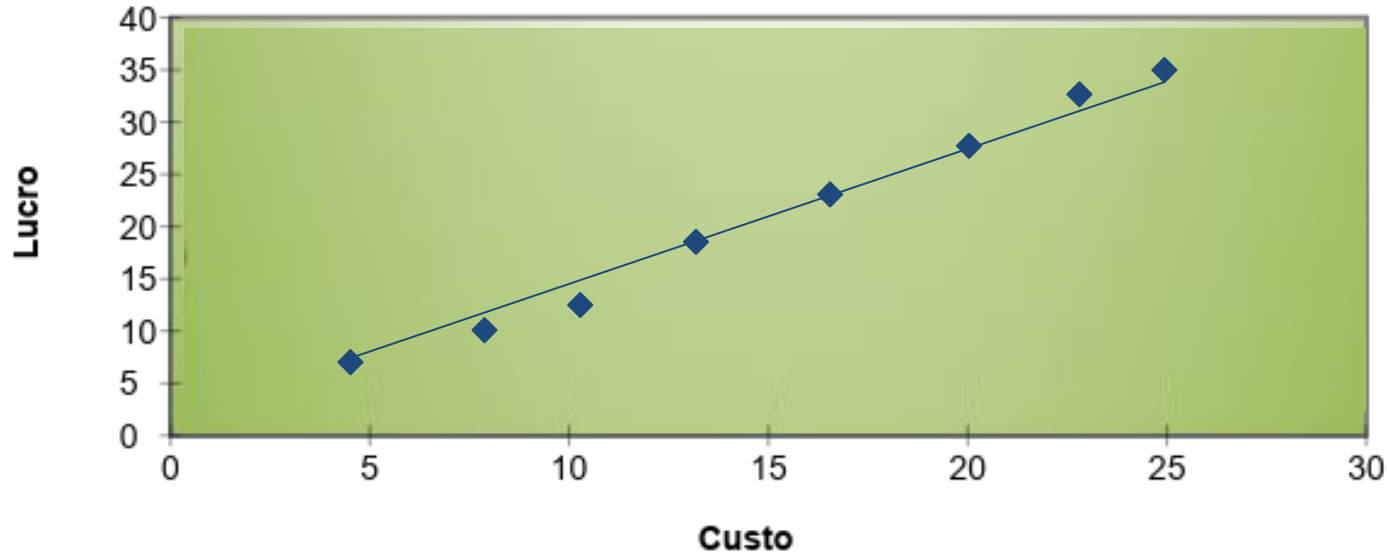
$r = -1$  existe correlação linear **negativa perfeita** entre as variáveis

---

$|r| \geq 0,70$  existe uma **forte** correlação linear entre as variáveis

$|r| < 0,70$  existe uma **fraca** correlação entre as variáveis

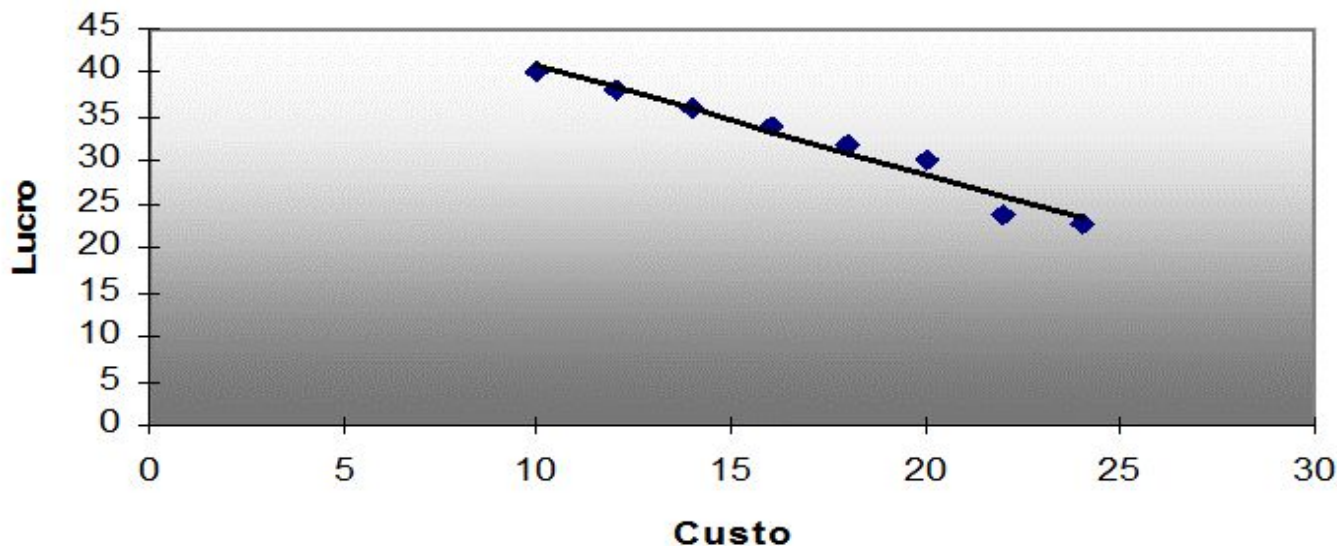
Coeficiente de correlação : 0,98



**Forte** Correlação Linear **Positiva** entre as Variáveis

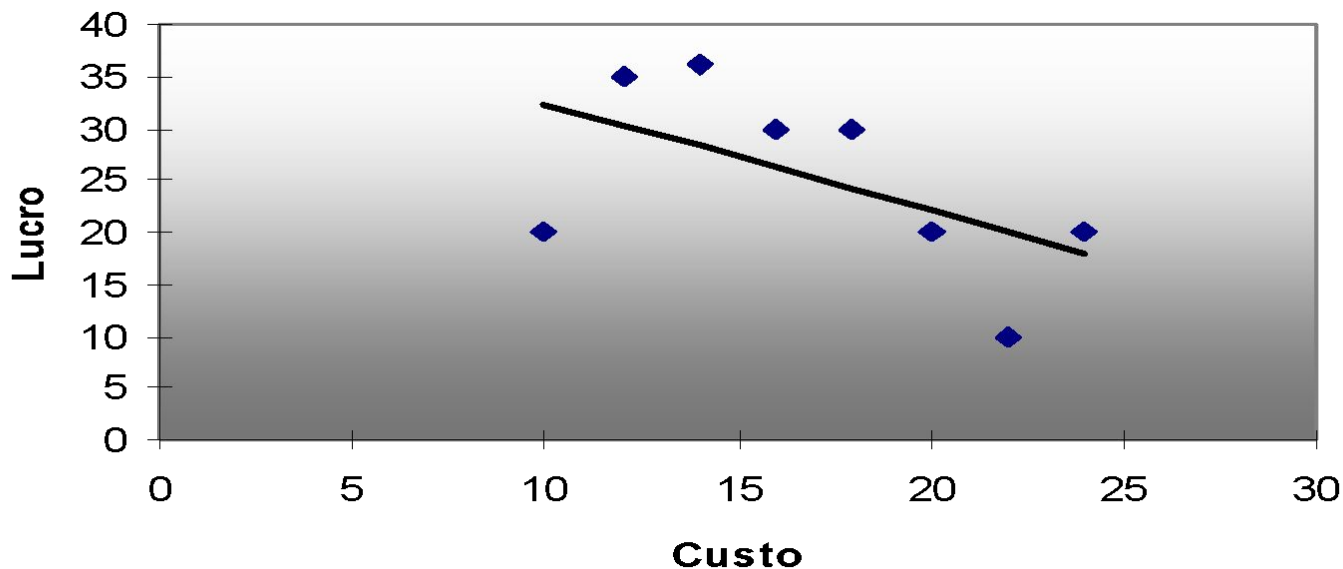


Coeficiente de correlação : -0,98



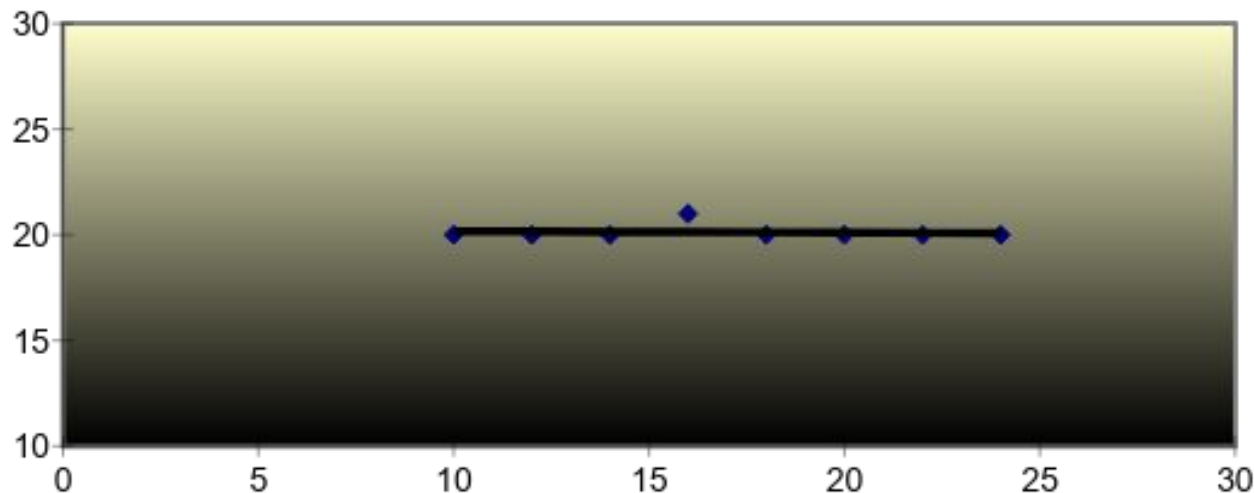
**Forte** Correlação Linear **Negativa** entre as Variáveis

Coeficiente de correlação : -0,55



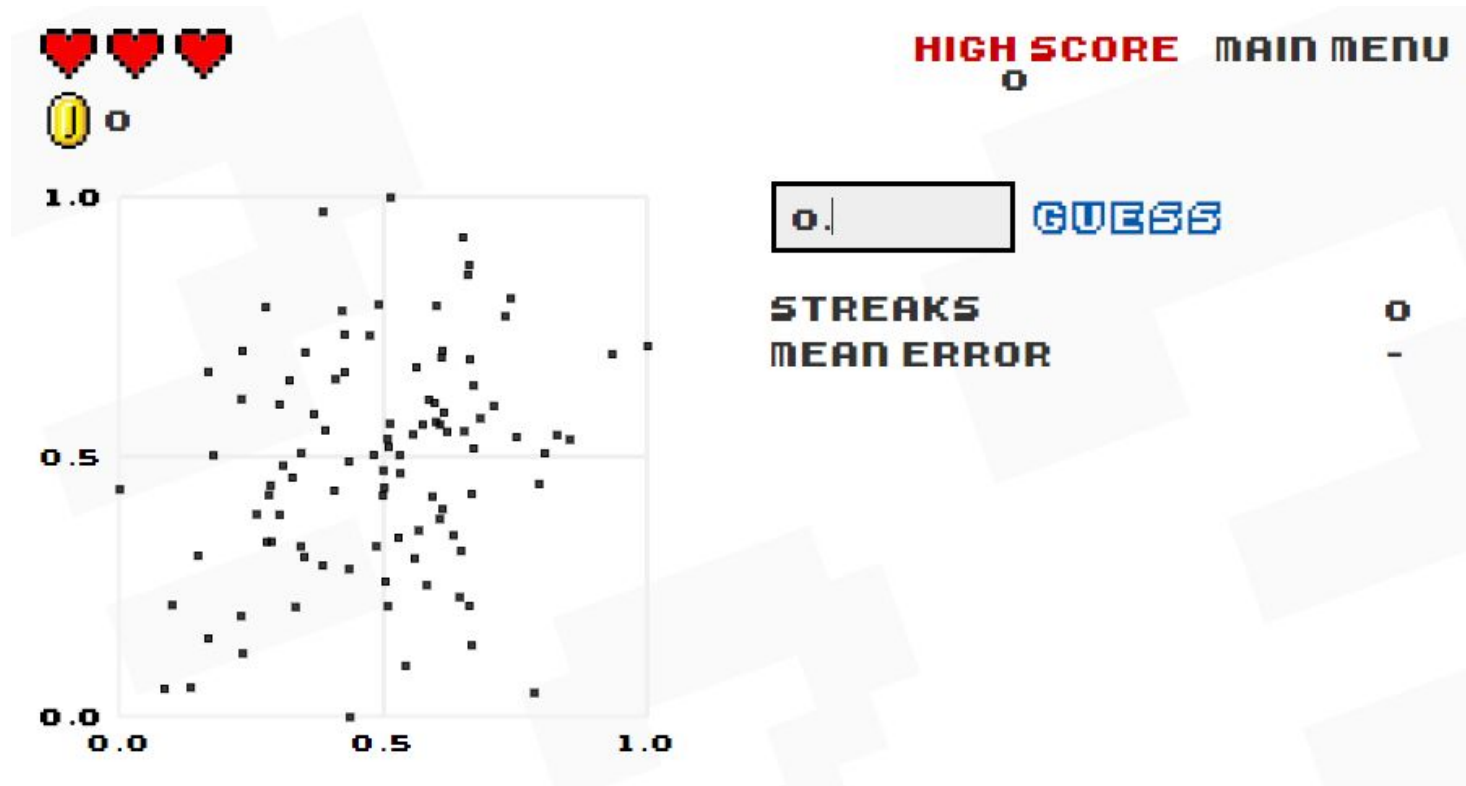
**Fraca** Correlação Linear **Negativa** entre as Variáveis

Coeficiente de correlação : -0,08



**Praticamente não existe** Correlação Linear entre as Variáveis

# Treine suas habilidades



Fonte: [Guess the Correlation](#)

# Case 1

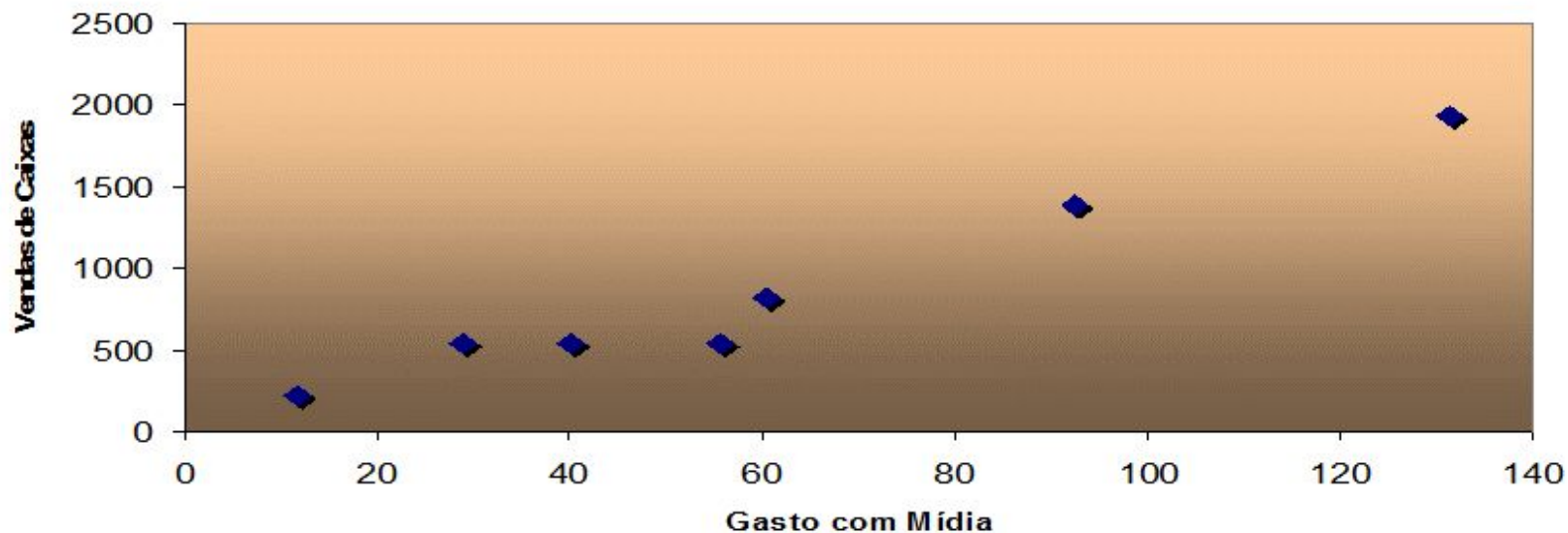
## Bebidas Refrigerantes

**Objetivo :** Verificar se há relação linear entre o gasto com mídia e as vendas de caixas

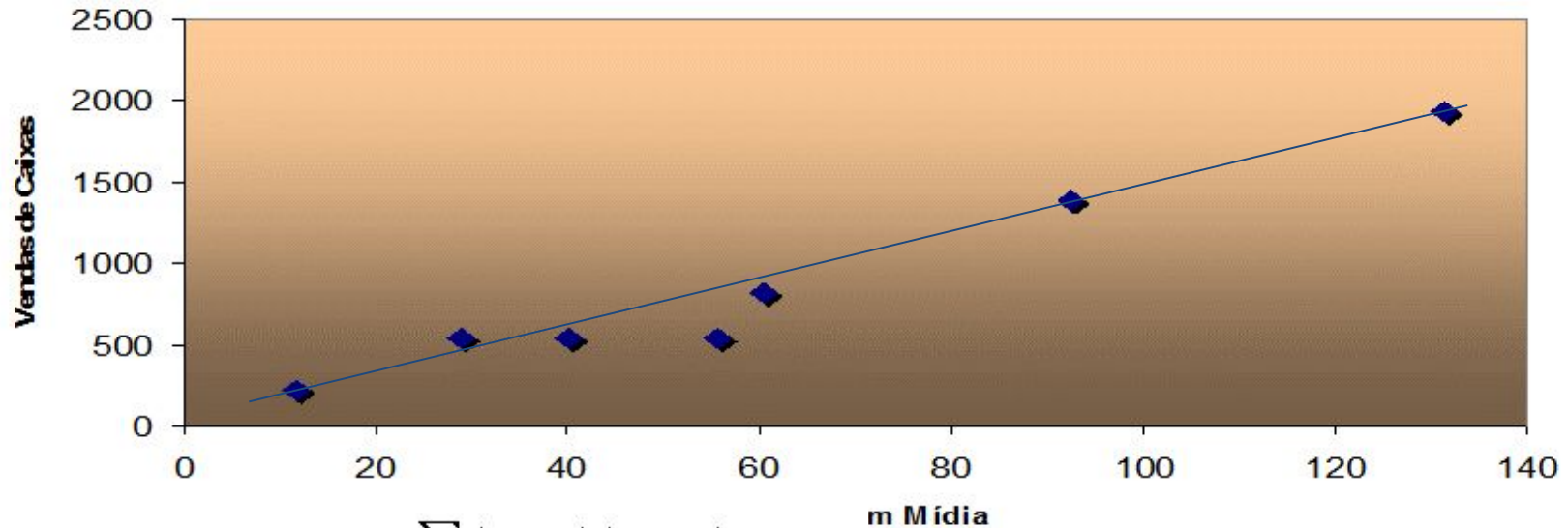
Marca	Gastos com Mídia (milhões de dólares)	Vendas de Caixas (milhões)
Coca-Cola	131.3	1929.2
Pepsi-Cola	92.4	1384.6
Coca-Cola Light	60.4	811.4
Sprite	55.7	541.5
Dr. Pepper	40.2	536.9
Mountain Dew	29.0	535.6
7-Up	11.6	219.5

Dados de 1997

## Vendas de Caixas x Gasto em Mídia



# Vendas de Caixas x Gasto em Mídia



$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right) \left(\sum (y_i - \bar{y})^2\right)}} \quad \xrightarrow{\text{m Mídia}} \quad r = 0,978$$

**Forte Correlação Linear Positiva entre as Variáveis**



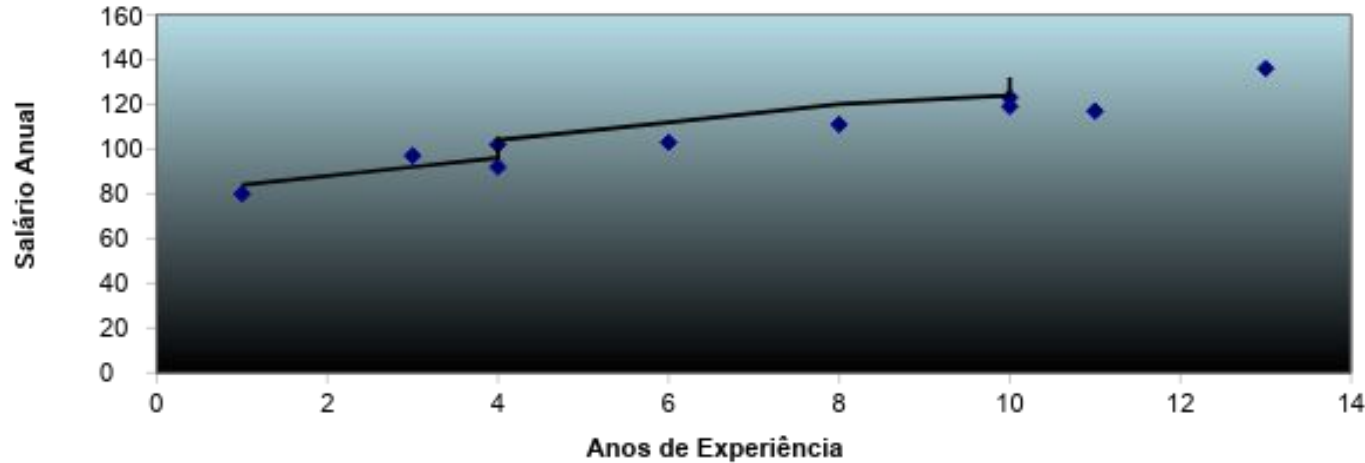
# Case 2

## Recursos Humanos

**Objetivo :** Estimar salário médio anual (Y) com base no tempo de experiência do funcionário (X)

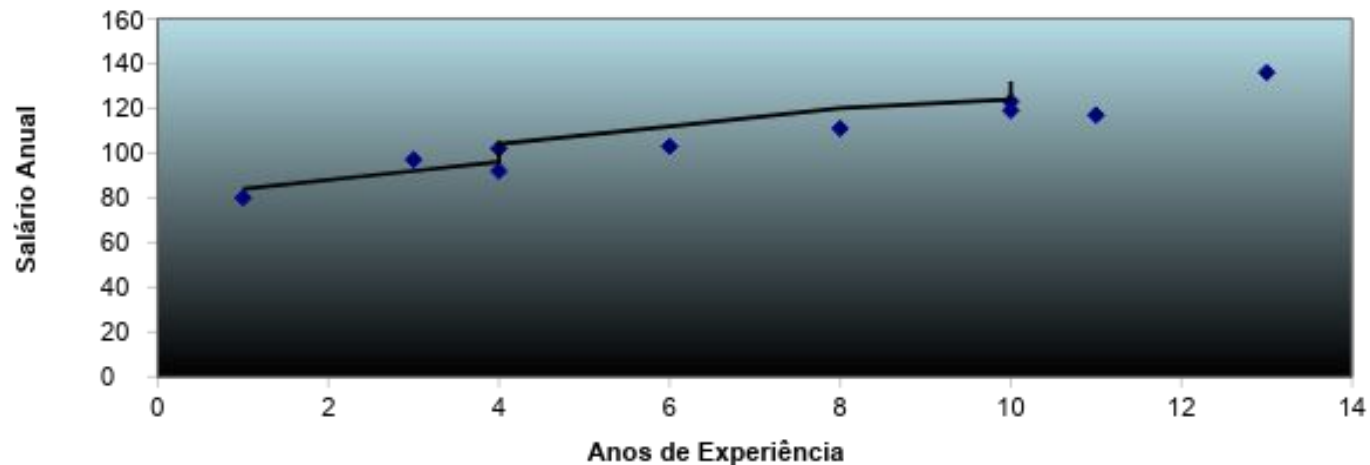
Funcionário	Tempo de Experiência (Anos) X	Salário Anual (R\$ 1.000) Y
1	1	80
2	3	97
3	4	92
4	4	102
5	6	103
6	8	111
7	10	119
8	10	123
9	11	117
10	13	136

## Salário anual (Y) vs. Tempo de experiência do funcionário (X)



Coeficiente de Correlação Linear :  $r = 0,964$

# Salário anual (Y) vs. Tempo de experiência do funcionário (X)



Reta de Regressão:

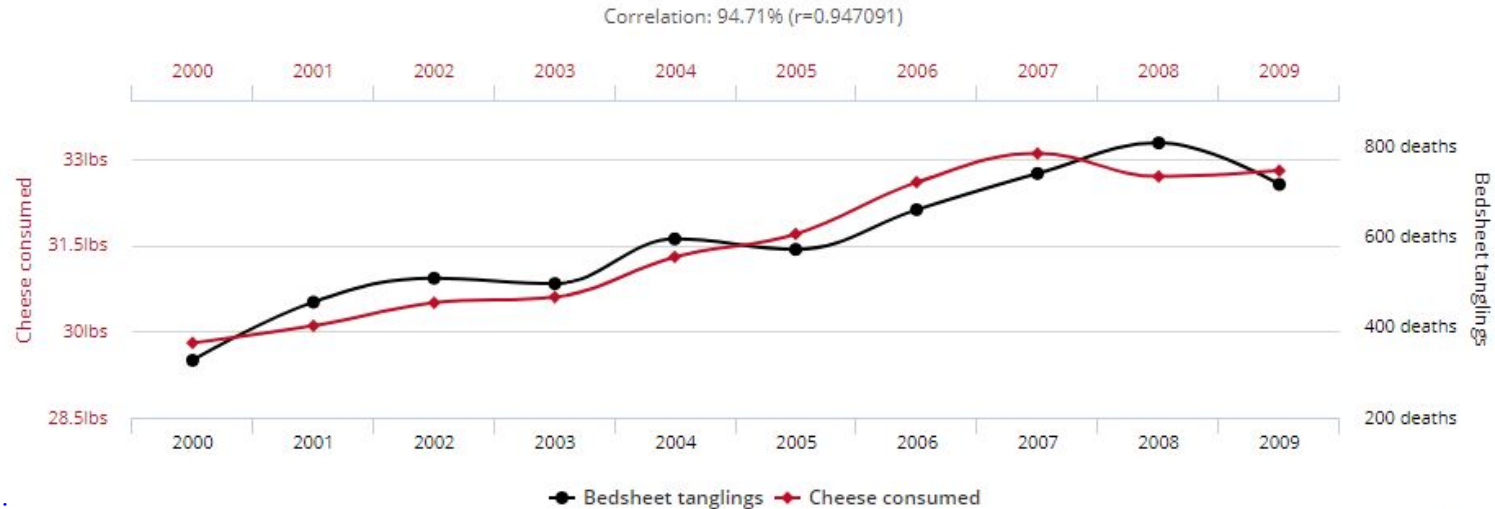
$$Y = \beta_0 + \beta_1 X$$

# Correlação não implica causalidade

## Consumo de queijo per capita

correlacionado com

## Número de pessoas que morreram enrolados em seus lençóis



Fonte: [Tylervigen](http://tylervigen.com)

tylervigen.com

# Equação do Modelo

(Reta da regressão linear)

# Modelo de Regressão Linear Simples

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \quad i=1,...,N$$

em que,

$Y_i$  : é o valor associado a i-ésima observação da variável resposta

$\beta_0$  e  $\beta_1$  : são parâmetros

$X_i$  : é o valor associado a i-ésima observação da variável explicativa

$\varepsilon_i$  : é o erro aleatório associado a i-ésima observação

$N$  : número de observações

# Modelo de Regressão Linear Simples

Equação da Reta:

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

intercepto

inclinação da reta

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

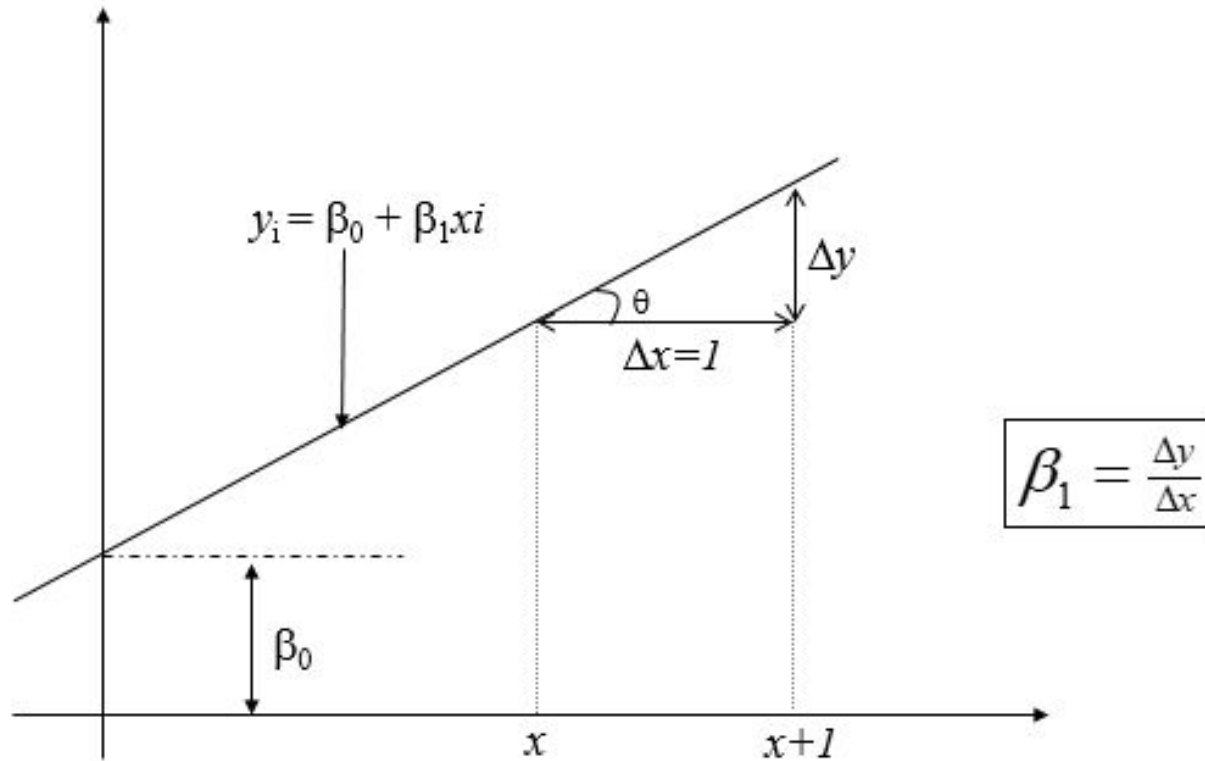
$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

$n$  : número de observações

em que  $\bar{X}$       média amostral da variável X  
 $\bar{Y}$       média amostral da variável Y



# Modelo de Regressão Linear Simples



# Suposições da regressão linear

As suposições para a regressão linear são:

**Linearidade:** a relação entre X e Y é linear;

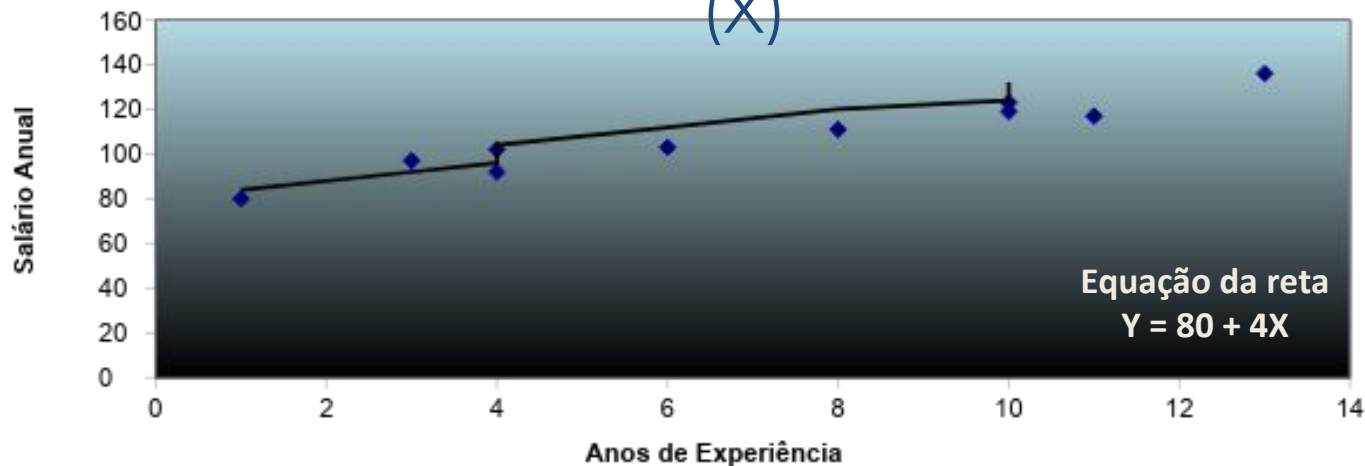
**Homocedasticidade:** a variância dos resíduos é a mesma para qualquer valor dado de X;

**Independência:** todas as observações são independentes umas das outras;

**Normalidade:** Y possui distribuição normal para qualquer valor dado de X.

*Obs: um erro é introduzido se a suposição básica sobre a linearidade do modelo não corresponder à realidade (tendência para a linearidade). Basicamente, o termo de erro deve ser aleatório e não depender de nenhuma entrada. Pode ser considerada a parte de Y que o modelo de regressão é incapaz de explicar.*

# Salário anual (Y) vs. Tempo de experiência do funcionário (X)



$$\beta_0 = 80$$

$$\beta_1 = 4$$

$$\text{Salário anual} = 80 + 4 (\text{anos de experiência})$$

Salário anual: variável dependente (Y)

Anos de experiência : variável independente (X)

# Interpretação dos Parâmetros

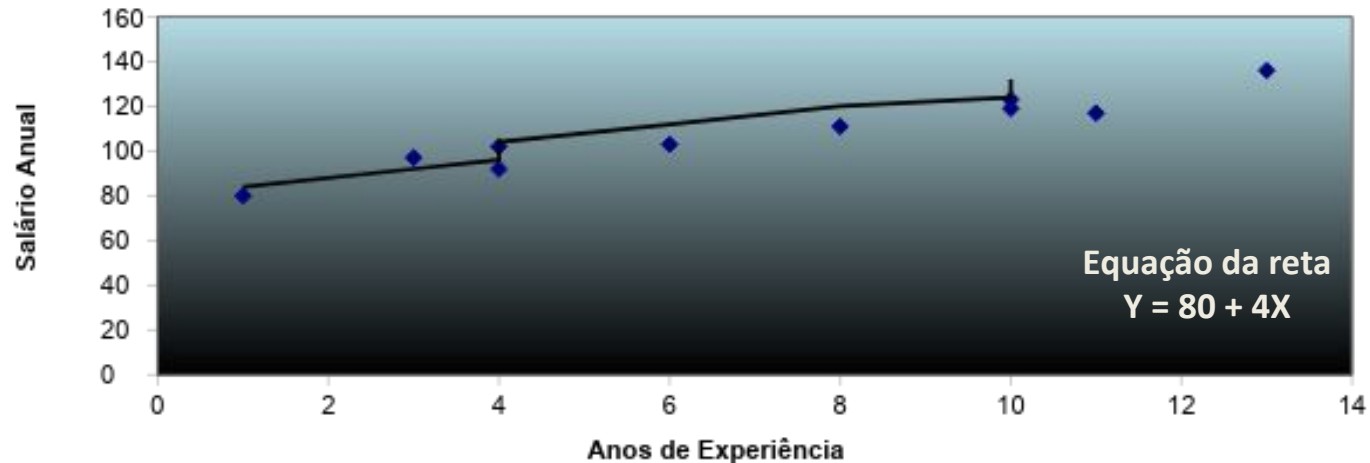
$$Y = \beta_0 + \beta_1 X$$

$$\text{Salário Anual} = 80 + 4 (\text{anos de experiência})$$

**80** : salário anual esperado para um funcionário que não possui um ano de experiência

**4** : acréscimo esperado na salário anual a cada variação de um ano no tempo de experiência do funcionário

## Salário anual vs. Tempo de experiência do funcionário



Salário anual =  $80 + 4$  (anos de experiência)

**Qual o salário anual estimado para um funcionário com 6 anos de experiência ?**

# Análise de Regressão Linear Simples - Excel

Salário anual (Y) vs. Tempo de experiência do funcionário (X)

## RESUMO DOS RESULTADOS

<i>Estatística de regressão</i>	
R múltiplo	0.964564633
R-Quadrado	0.93038493
R-quadrado ajustado	0.921683047
Erro padrão	4.609772229
Observações	10

### Coeficiente de determinação

93,04% da variabilidade da venda anual é explicada pelo tempo de experiência

## ANOVA

	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
Regressão	1	2272	2272	106.9176	6.60903E-06
Resíduo	8	170	21.25		
Total	9	2442			

	<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
Interseção	80	3.075344937	26.01334	5.12E-09	72.90823727	87.09176273
Tempo de Experiência	4	0.386843492	10.3401	6.61E-06	3.107936731	4.892063269

Parâmetros

Probabilidades do teste de hipótese

Funcionário	Tempo de Experiência (Anos) X	Salário anual (R\$ 1.000) Y observado	Salário Anual estimado pelo modelo Y ajustado	Resíduo
1	1	80	84	-4
2	3	97	92	5
3	4	92	96	-4
4	4	102	96	6
5	6	103	104	-1
6	8	111	112	-1
7	10	119	120	-1
8	10	123	120	3
9	11	117	124	-7
10	13	136	132	4

$\varepsilon_i$ : erro aleatório = resíduo = Y observado – Y ajustado

# Análise de Regressão Linear Simples - R





# CARS

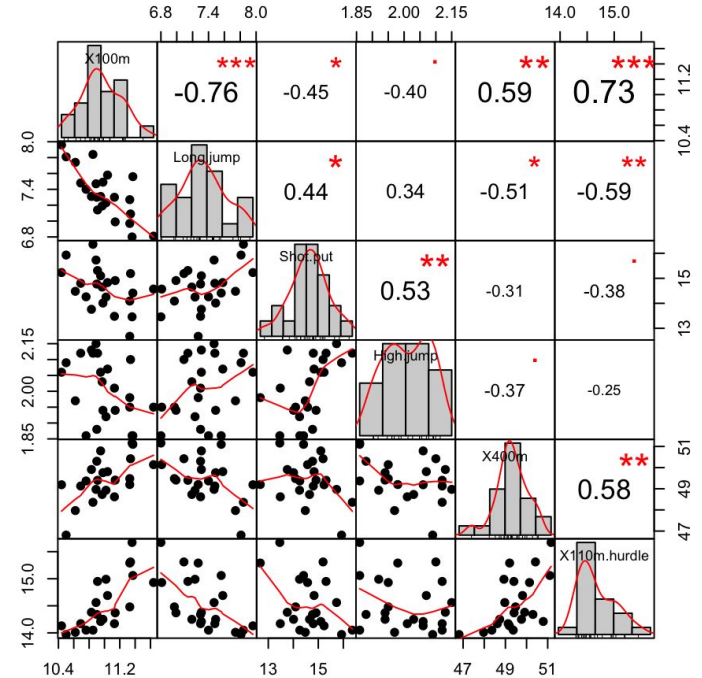
Vamos utilizar o conjunto de dados `cars`. Acesse os dados digitando `cars` no seu console do R. Ao todo são 50 observações (linhas) e 2 variáveis (colunas) - *dist* e *speed*.

Vamos imprimir as primeiras seis observações:

```
head(cars)  # Exibir as 6 primeiras observações
#>   speed dist
#> 1     4    2
#> 2     4   10
#> 3     7    4
#> 4     7   22
#> 5     8   16
#> 6     9   10
```

# Análises Gráficas

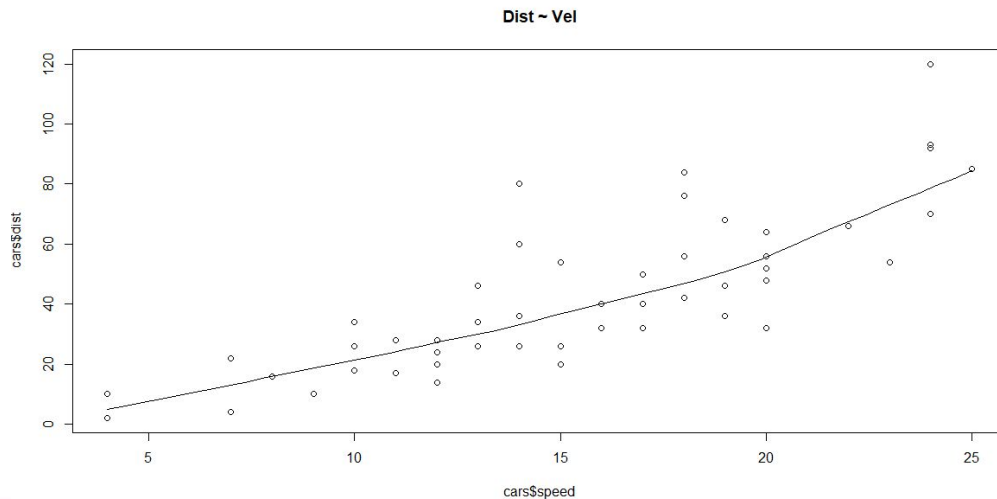
O objetivo desta análise é construir um modelo de regressão simples que podemos usar para prever a distância (dist), estabelecendo uma relação linear estatisticamente significativa com a velocidade (speed). Mas antes de entrar no modelo, vamos tentar entender essas variáveis graficamente:



# Gráficos de Dispersão - *Scatterplot*

O scatterplot pode nos ajudar a visualizar qualquer relação linear entre a variável dependente e a independente

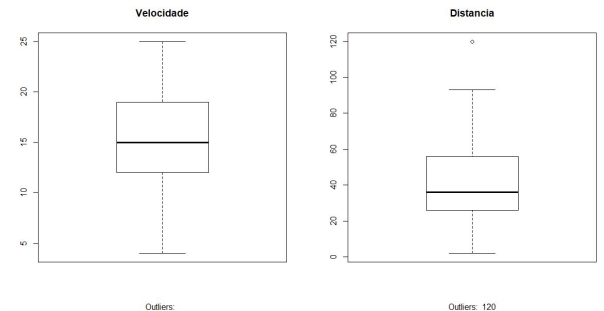
```
scatter.smooth(x=cars$speed, y=cars$dist, main="Dist ~ Vel.") # scatterplot
```



# BoxPlot - Verificar Outlier

Geralmente, qualquer ponto de dados que esteja fora do  $QI - 1,5 * (1,5 * IQR)$  é considerado um valor *outlier*, em que o IQR é calculado como a distância entre os valores do percentil 25 e 75 para essa variável.

```
par(mfrow=c(1, 2)) # Dividir a área de gráfico em 2
boxplot(cars$speed, main="Velocidade", sub=paste("Outliers: ",
boxplot.stats(cars$speed)$out)) # box plot para a variável 'speed'
boxplot(cars$dist, main="Distancia", sub=paste("Outliers: ",
boxplot.stats(cars$dist)$out)) # box plot para 'distância'
```



# Gráfico de Normalidade

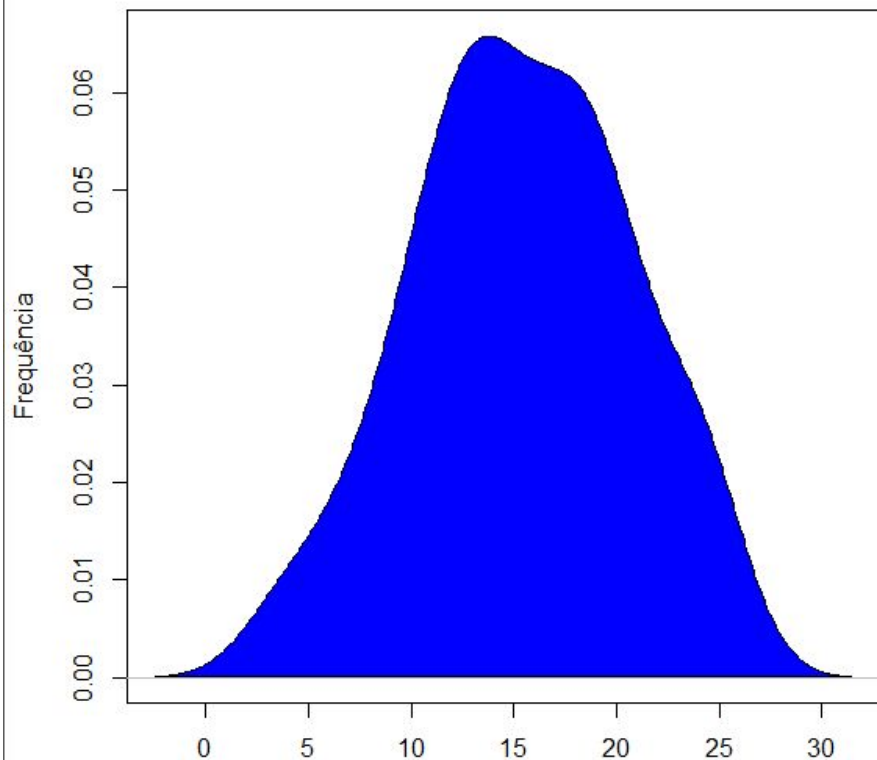
Verifique se a variável de resposta está próxima da normalidade

```
library(e1071)

par(mfrow=c(1, 2)) # Dividir a área de gráfico em 2
plot(density(cars$speed),
     main="Gráfico de densidade: Velocidade",
     ylab="Frequência",
     sub=paste("Skewness:",
               round(e1071::skewness(cars$speed), 2))) # gráfico de densidade para 'speed'
polygon(density(cars$speed), col="red")

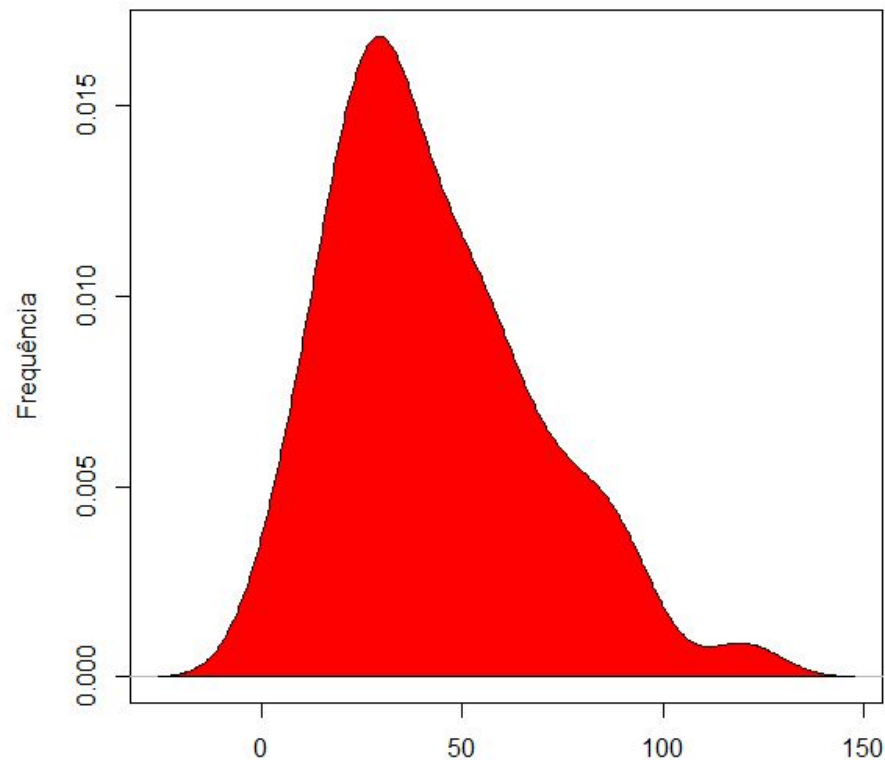
plot(density(cars$dist), main="Gráfico de densidade: Distância",
     ylab="Frequência",
     sub=paste("Skewness:", round(e1071::skewness(cars$dist), 2))) # Densidade para 'dist'
polygon(density(cars$dist), col="red")
```

Gráfico de densidade: Velocidade



N = 50 Bandwidth = 2.15  
Skewness: -0.11

Gráfico de densidade: Distância

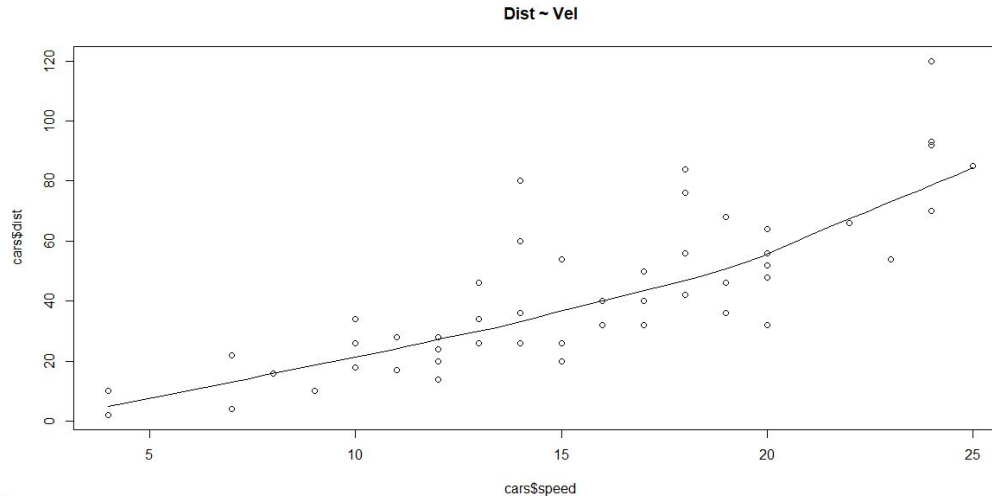


N = 50 Bandwidth = 9.214  
Skewness: 0.76

# Correlação

Se observarmos no *Scatterplot*, para cada instância onde a velocidade aumenta, a distância também aumenta junto com ela, então há uma alta correlação positiva entre eles e, portanto, é esperado que a correlação entre as variáveis esteja próxima de 1.

```
cor(cars$speed, cars$dist)  # Calcular a correlação entre velocidade e distância  
#> [1] 0.8068949
```



# Modelo Linear

No R, a função utilizada para construir o modelo linear é `lm()`. A função `lm()` recebe dois argumentos principais: 1. Fórmula, 2. Dados. Normalmente os dados vem de um `data.frame` e a fórmula é um objeto de classe fórmula.

```
linearMod <- lm(dist ~ speed, data=cars) # modelo de regressão linear
print(linearMod)

#> Call:
#> lm(formula = dist ~ speed, data = cars)
#>
#> Coefficients:
#> (Intercept)      speed
#>    -17.579      3.932
```



# Modelo Linear

No R, a função utilizada para construir o modelo linear é `lm()`. A função `lm()` recebe dois argumentos principais: 1. Fórmula, 2. Dados. Normalmente os dados vem de um `data.frame` e a fórmula é um objeto de classe fórmula.

```
linearMod <- lm(dist ~ speed, data=cars) # modelo de regressão linear
print(linearMod)

#> Call:
#> lm(formula = dist ~ speed, data = cars)
#>
#> Coefficients:
#> (Intercept)      speed
#>    -17.579      3.932
```

Podemos notar na saída que a parte dos 'Coeficientes' possuem dois componentes: Intercepto: -17.579, velocidade: 3.932.

# Modelo Linear

No R, a função utilizada para construir o modelo linear é `lm()`. A função `lm()` recebe dois argumentos principais: 1. Fórmula, 2. Dados. Normalmente os dados vem de um `data.frame` e a fórmula é um objeto de classe fórmula.

```
linearMod <- lm(dist ~ speed, data=cars) # modelo de regressão linear
print(linearMod)

#> Call:
#> lm(formula = dist ~ speed, data = cars)
#>
#> Coefficients:
#> (Intercept)      speed
#>    -17.579      3.932
```

Podemos notar na saída que a parte dos 'Coeficientes' possuem dois componentes: Intercepto: -17.579, velocidade: 3.932. Ou seja:

## Equação do Modelo

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

$$\text{Dist} = -17.579 + 3.932 * \text{Vel}$$

Já podemos utilizar nosso modelo?

# Diagnósticos Modelo Linear

```
summary(linearMod)  # Resumo do modelo
```

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min       1Q   Median       3Q      Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

# Interpretando o Output - R

1

```
Residuals:
  Min       1Q   Median       3Q      Max
-123.68  -26.78   -5.07   24.66   95.04
```

**Residuais:** Os resíduos são a diferença entre os valores reais da variável que você prevê e os valores previstos da sua regressão  $y - \hat{y}$ .

Para a maioria das regressões, você deseja que seus resíduos pareçam uma distribuição normal quando plotados. Se nossos resíduos são normalmente distribuídos, isso indica a média da diferença entre nossas previsões e os valores reais são próximos de 0 (bom).

# Interpretando o Output - R

Coefficients:	2	3	4	5
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.212e+02	9.352e+00	45.03	<2e-16 ***
Renda	5.348e-02	2.766e-03	19.33	<2e-16 ***
---				

2 - **Coefficiente de estimação:** O coeficiente estimado é o valor da inclinação calculada pela regressão.

3 - **Erro padrão da regressão:** Mensura a variabilidade na estimativa do coeficiente. É a raiz quadrada da variância estimada dos resíduos e indica o grau de dispersão dos erros de previsão dentro da amostra na hipótese de normalidade.

4 - **t-valor do coeficiente estimado:** Podemos interpretar o valor de t assim: Um valor t maior indica que é menos provável que o coeficiente não seja igual a zero apenas por acaso. Então, quanto maior o valor de t, melhor.

5 - **p-valor** (probabilidade de significância): a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, sob a hipótese nula

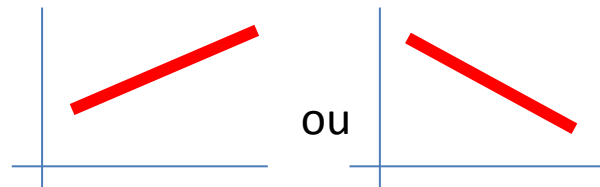
# Hipótese de Interesse

inclinação da reta

$H_0: \beta_1 = 0$ ; não existe relação linear entre as variáveis



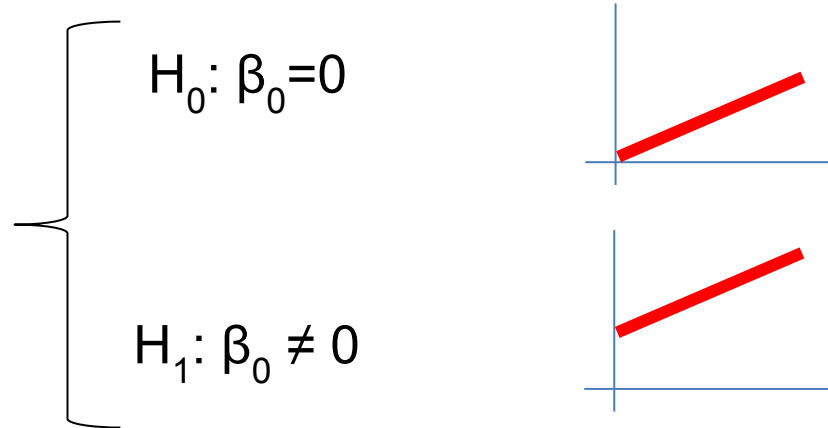
$H_1: \beta_1 \neq 0$ ; existe relação linear entre as variáveis



	<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
Interseção	80	3,075344937	26,01334	5,12002E-09	72,90823727	87,09176273
Tempo de Experiência	4	0,386843492	10,3401	6,60903E-06	3,107936731	4,892063269

# Hipótese de Interesse

intercepto



	<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
Interseção	80	3,075344937	26,01334	5,12002E-09	72,90823727	87,09176273
Tempo de Experiência	4	0,386843492	10,3401	6,60903E-06	3,107936731	4,892063269



# Calculando Teste t e p-valor

Quando os coeficientes do modelo e o erro padrão são conhecidos, a fórmula para calcular a estatística t e o valor-p é a seguinte:

$$t - \text{Statistic} = \frac{\beta - \text{coefficient}}{\text{Std. Error}}$$

```
modelSummary <- summary(linearMod)
modelCoeffs <- modelSummary$coefficients # Coeficientes
beta.estimate <- modelCoeffs["speed", "Estimate"] # estimativa de beta para velocidade
std.error <- modelCoeffs["speed", "Std. Error"] # std.error para velocidade
t_value <- beta.estimate/std.error # calcular teste t
p_value <- 2*pt(-abs(t_value), df=nrow(cars)-ncol(cars)) # calcular p Valor
f_statistic <- linearMod$fstatistic[1] # estatística de F
f <- summary(linearMod)$fstatistic # parâmetros para o cálculo do modelo
model_p <- pf(f[1], f[2], f[3], lower=FALSE)
```

# Interpretando o Output - R

```
>>>summary(fit)
```

```
Residuals:
    Min       1Q   Median       3Q      Max    1
-123.68  -26.78   -5.07   24.66   95.04

Coefficients:    2        3        4        5
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.212e+02  9.352e+00  45.03  <2e-16 ***
Renda       5.348e-02  2.766e-03  19.33  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.93 on 103 degrees of freedom
Multiple R-squared:  0.784,    Adjusted R-squared:  0.7819
F-statistic: 373.8 on 1 and 103 DF,  p-value: < 2.2e-16
```

**Coefficiente de determinação**  
78,4% da variabilidade da venda  
anual é explicada pelo tempo de  
experiência

6

7

Fonte: [lm-summary\(yhat\)](#)

# Interpretando o Output - R

```
Residual standard error: 38.93 on 103 degrees of freedom  
Multiple R-squared: 0.784, Adjusted R-squared: 0.7819  
F-statistic: 373.8 on 1 and 103 DF, p-value: < 2.2e-16
```

6  
7

**Erro padrão da regressão:** é a raiz quadrada da variância estimada dos resíduos e indica o grau de dispersão dos erros de previsão dentro da amostra na hipótese de normalidade.

**R<sup>2</sup>:** é uma medida do grau de proximidade entre os valores estimados e observados da variável dependente dentro da amostra utilizada para estimar a regressão, sendo portanto uma medida do sucesso da estimativa.

**R<sup>2</sup> Adjusted:** é uma medida semelhante ao R-quadrado mas que, ao contrário deste, não aumenta com a inclusão de variáveis independentes não significativas.

# R-Quadrado

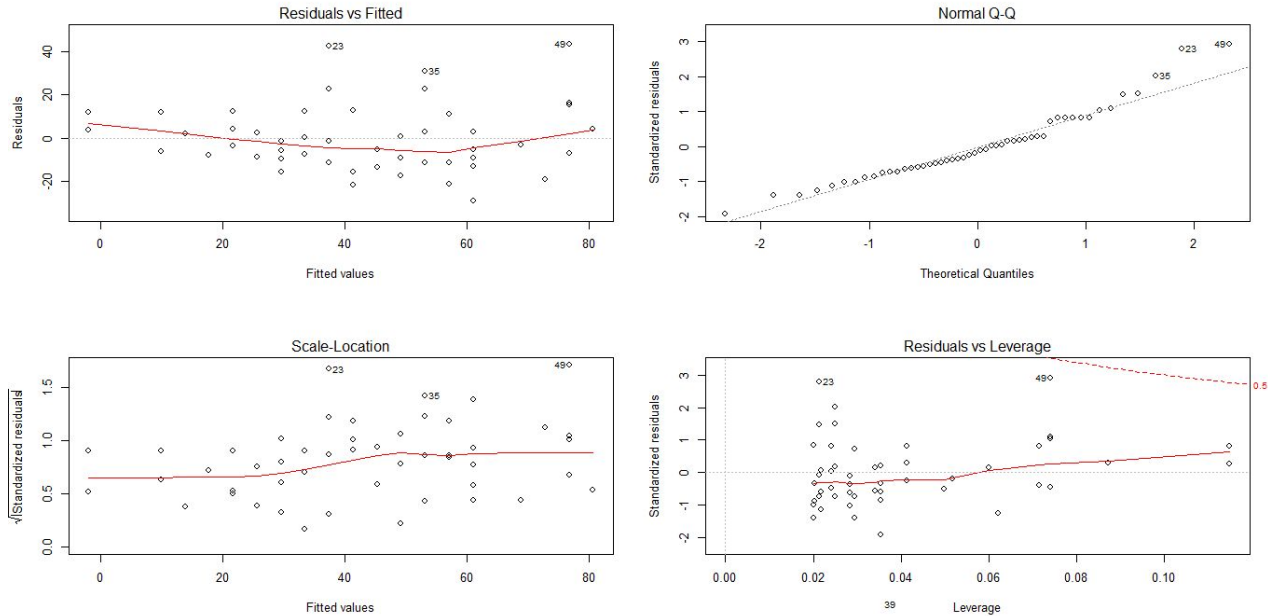
Basicamente, o que o R-Quadrado nos diz é a proporção da variação na variável dependente (resposta) que foi explicada por este modelo.

$$R^2 = 1 - \frac{SSE}{SST}$$

$$SSE = \sum_i^n (y_i - \hat{y}_i)^2 \text{ and } SST = \sum_i^n (y_i - \bar{y}_i)^2$$

# Análise Gráfica

```
par(mfrow=c(2,2)) #Criar matriz com duas linhas e duas colunas  
plot(linearMod) #plotar gráfico
```



# Principais Métricas

Estatística	Critério
R-Quadrado	Maior melhor ( $>0,7$ )
R-Quadrado Ajustado	Maior Melhor
Teste-F	Maior melhor
Std. Error	Próximo de 0 melhor
Teste. T	Deve ser maior que 1,96 para o p-valor ser menor que 0,05
MAPE (Erro percentual absoluto médio)	Menor melhor
MSE (Erro quadrático médio)	Menor melhor

# Análise de Regressão Linear Múltipla

# Modelo de Regressão Linear Múltipla - Teórico

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \varepsilon$$

$Y$ : variável dependente

$X_1, \dots, X_n$ : variáveis independentes

$\beta_0$  e  $\beta_1$ : são parâmetros

$\varepsilon$ : erro aleatório associado ao modelo



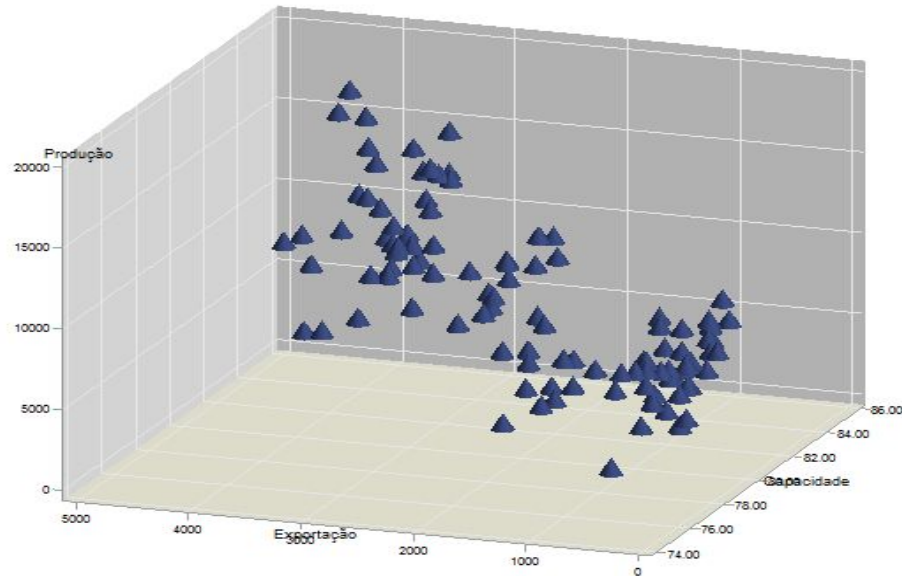
# Modelo de Regressão Linear Múltipla

Tem o objetivo de projetar uma variável de interesse em função de várias variáveis auxiliares.

Por exemplo projetar o número de clientes que irão adquirir um cartão de crédito em função do número de benefícios oferecidos.

# Representação Gráfica

# Regressão Linear Múltipla



# Case 3

## Marketing

## Objetivo : Estimar o faturamento (Y) com base no investimento em anúncios ( $X_1$ e $X_2$ )

Anúncio de Televisão (R\$ 1.000) $X_1$	Anúncio de Jornal (R\$ 1.000) $X_2$	Faturamento Bruto Semanal (R\$ 1.000) Y
5,0	1,5	96
2,0	2,0	90
4,0	1,5	95
2,5	2,5	92
3,0	3,3	95
3,5	2,3	94
2,5	4,2	94
3,0	2,5	94

# Resumo dos Resultados no Excel

<i>Estatística de regressão</i>	
R múltiplo	0,958663444
R-Quadrado	0,9190356
R-quadrado ajustado	0,88664984
Erro padrão	0,642587303
Observações	8

É  $< 0,10 \rightarrow$  existe relação linear entre as variáveis

<i>ANOVA</i>					
	<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significação</i>
Regressão	2	23,43540779	11,7177039	28,37776839	0,001865242
Resíduo	5	2,064592208	0,412918442		
Total	7	25,5			

	<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>	<i>95% superiores</i>
Interseção	<b>83,230092</b>	1,573868952	52,88247894	<b>0,000000</b>	79,18433275	87,27585063
Anuncio de Televisao	<b>2,2901836</b>	0,304064556	7,531899313	<b>0,000653</b>	1,508560796	3,071806446
Anuncio de Jomal	<b>1,3009891</b>	0,320701597	4,056696662	<b>0,009761</b>	0,476599398	2,125378798

# Resumo dos Resultados

Call:

```
lm(formula = "Faturamento ~ AnuncioTelevisao + AnuncioJornal",  
    data = vendas)
```

Residuals:

1	2	3	4	5	6	7	8
-0.6325	-0.4124	0.6577	-0.2080	0.6061	-0.2380	-0.4197	0.6469

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	83.2301	1.5739	52.882	4.57e-08	***
AnuncioTelevisao	2.2902	0.3041	7.532	0.000653	***
AnuncioJornal	1.3010	0.3207	4.057	0.009761	**

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6426 on 5 degrees of freedom

Multiple R-squared: 0.919, Adjusted R-squared: 0.8866

F-statistic: 28.38 on 2 and 5 DF, p-value: 0.001865

# Equação de Projeção

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\text{Fatur} = 83,23 + 2,29 * \text{tv} + 1,30 * \text{jornal}$$



# R-Quadrado Ajustado

À medida que vamos adicionando mais variáveis (X) ao nosso modelo, o valor do R-Quadrado do novo modelo sempre será maior que o do subconjunto menor.

Então o R-Quadrado Ajustado “penaliza” o valor total do número de termos (preditores de leitura) em seu modelo. Portanto, ao comparar modelos, é uma boa prática observar o valor R-Quadrado Ajustado ao R-quadrado.

$$R^2_{adj} = 1 - \frac{MSE}{MST}$$

$$MSE = \frac{SSE}{(n-q)}$$

$$MST = \frac{SST}{(n-1)}$$

$n$  é o número de observações e o  $q$  é o número de coeficientes no modelo

# R-Quadrado Ajustado

Portanto, movendo-se em torno dos numeradores e denominadores, a relação entre  $R^2$  e  $R^2$ -adj torna-se:

$$R^2_{adj} = 1 - \left( \frac{(1 - R^2)(n - 1)}{n - q} \right)$$

# Teste F

O teste F valida se ao menos uma variável explica o modelo

# Test F | Hipótese de Interesse

$$\begin{cases} H_0: \beta_1 = \beta_2 = \dots = \beta_n = 0 \\ H_1: \beta_1 \neq 0 \text{ ou } \beta_2 \neq 0 \text{ ou } \dots \text{ ou } \beta_n \neq 0 \end{cases}$$

10	ANOVA					
11		<i>gl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>F de significância</i>
12	Regressão	2	14243,09333	7121,547	9,4E+29	2,01E-34
13	Resíduo	3	2,27294E-26	7,58E-27		
14	Total	5	14243,09333			
15						
16		<i>Coeficientes</i>	<i>Erro padrão</i>	<i>Stat t</i>	<i>valor-P</i>	<i>95% inferiores</i>
17	Interseção	0,5	1,08879E-13	4,59E+12	2,28E-38	0,5
18	Variável X 1	2	4,95082E-14	4,04E+13	3,35E-41	2
19	Variável X 2	1,2	4,45143E-15	2,7E+14	1,13E-43	1,2
20						

Quando este número for < 0,10  
**existe** relação linear entre as  
variáveis

# Análise de Regressão Linear Múltipla



# The Boston Housing Dataset

O conjunto de dados contém informações coletadas pelo Serviço de Censo dos EUA sobre habitação na área de Boston.

Os dados foram originalmente publicados por Harrison, D. e Rubinfeld, D.L. "*Hedonic prices and the demand for clean air*", J. Environ. Economics & Management, volume 5, 81-102, 1978. [Link para a publicação](#)

O trabalho investiga os problemas metodológicos associados ao uso de dados do mercado habitacional para medir a disposição de pagar por ar limpo.



# Dataset

O Dataset disponibilizado contém 506 observações com 14 variáveis, sendo:

1. **CRIM** - taxa de criminalidade per capita por cidade
2. **ZN** - proporção de terrenos residenciais destinados a lotes com mais de 25.000 pés quadrados (25.000 pés<sup>2</sup> = 2.322,576 mts<sup>2</sup>)
3. **INDUS** - proporção de hectares comerciais não varejistas por cidade.
4. **CHAS** - Variável dummy de Charles River (1 se o setor delimita rio; 0 caso contrário)
5. **NOX** - concentração de óxidos nítricos (partes por 10 milhões)
6. **RM** - número médio de quartos por habitação
7. **AGE** - proporção de unidades ocupadas pelo proprietário construídas antes de 1940
8. **DIS** - distâncias ponderadas para cinco centros de emprego de Boston
9. **RAD** - índice de acessibilidade para rodovias radiais
10. **TAX** - taxa de imposto sobre propriedades de valor integral por US \$ 10.000
11. **PTRATIO** - proporção aluno-professor por cidade
12. **B - 1000(Bk - 0.63)^2** onde Bk é a proporção de negros por cidade
13. **LSTAT** - % *lower status* da população
14. **MEDV** - Valor mediano de residências ocupadas pelo proprietário em US \$ 1.000

# Utilizar a folha de exercícios Boston Housing



# Referências Bibliográficas

Anderson, R. A., Sweeney, J. D. e Williams, T. A. Estatística Aplicada à Administração e Economia. Pioneira. Thomson Learning. 2003

Statistics for Business By Robert Stine, Dean Foster

An Introduction to Statistical Learning, with Application in R. By G. Casella, S. Fienberg, I. Olkin

