

# **Análise de Regressão: Teoria e Prática**

**Departamento de Estatística e Matemática Aplicada**

Rafael Braz, Ronald Targino, Juvêncio Nobre e Manoel Santos-Neto

2025-08-18

# Índice

# Prefácio

Este livro....

# 1 Introdução

## 2 Preliminares

## 3 Análise de Regressão

### 3.1 Regressão Linear Simples

#### 3.1.1 Motivação

**Regressão Linear Simples:** É um método estatístico que nos permite resumir e estudar as relações entre duas variáveis quantitativas:

- Uma variável, denotada por  $x$ , é considerada como preditora, explicativa ou variável independentes.
- A outra variável, denotada por  $y$ , é considerada como a resposta, resultado ou variável dependente.

Usaremos os termos “**preditor**” e “**resposta**” para nos referirmos às variáveis utilizadas neste curso. Os outros termos são mencionados apenas para torná-lo ciente deles caso você os encontre em outros materiais. A regressão linear simples recebe o adjetivo “*simples*”, porque diz respeito ao estudo de apenas uma variável preditora. Em contraste, a regressão linear múltipla, que estudaremos mais adiante neste curso, recebe o adjetivo “*múltipla*”, porque diz respeito ao estudo de duas ou mais variáveis preditoras.

No slide anterior, foi possível observar que se você conhece a temperatura em graus Celsius, pode usar uma equação para determinar exatamente a temperatura em graus Fahrenheit.

Agora serão apresentadas outros exemplos de relações determinísticas.

1. Circunferência =  $\pi \times$  diâmetro.
2. **Lei de Hooke:**  $Y = \alpha + \beta X$ , em que  $Y$  é a quantidade de alojamento em uma mola e  $X$  é o peso aplicado.
3. **Lei de Ohm:**  $I = V/r$ , em que  $V$  é a tensão aplicada,  $r$  é a resistência elétrica e  $I$  é a corrente elétrica.
4. **Lei de Boyle:** Para uma temperatura constante,  $P = \alpha/V$ , em que  $P$  é a pressão,  $\alpha$  é uma constante para cada gás e  $V$  é o volume do gás.

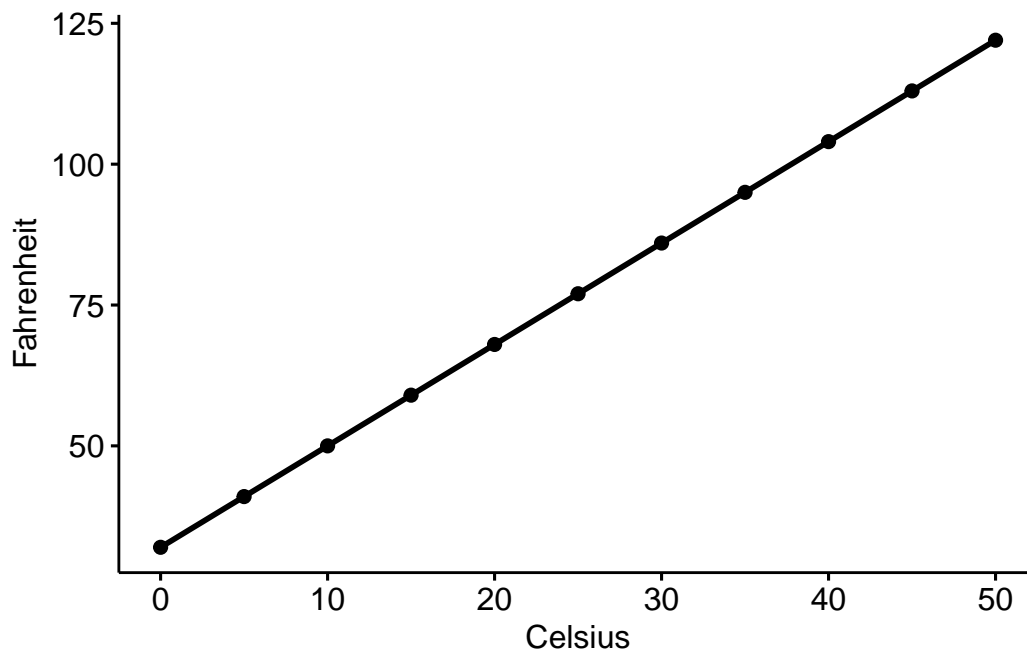
Para cada uma dessas relações determinísticas, a equação descreve exatamente a relação entre as duas variáveis. Esta disciplina não examina relacionamentos determinísticos. Em vez disso, estamos interessados em relações estatísticas, nas quais a relação entre as variáveis não é perfeita.

Primeiro devemos deixar claro quais tipos de relacionamentos não estudaremos neste curso, ou seja, relacionamentos determinísticos (ou funcionais). Abaixo está um exemplo de uma relação determinística.

```
library(ggpubr)
```

Carregando pacotes exigidos: ggplot2

```
cels <- seq(0, 50, by = 5)
fahr <- (9/5)*cels + 32
data <- data.frame(x = cels, y = fahr)
ggscatter(data,
  x = "x",
  y = "y",
  xlab = "Celsius",
  ylab = "Fahrenheit",
  add = "reg.line")
```



Observe que os pontos de dados observados caem diretamente em uma linha. Como você deve se lembrar, a relação entre graus Fahrenheit e graus Celsius é conhecida como:

$$\text{Fahrenheit} = (9/5) \times \text{Celsius} + 32.$$

Agora iremos apresentar um exemplo de relação estatística. A variável resposta  $Y$  é a mortalidade por câncer de pele (por 10 milhões de pessoas) e a variável preditora  $X$  é a latitude no centro de cada um dos 48 estados americanos (dados de câncer de pele dos EUA). Os dados foram obtidos na década de 1950, então o Alasca e o Havaí ainda não eram estados. Além disso, Washington, DC está incluído no conjunto de dados, embora não seja tecnicamente um estado.

```
library(tidyverse)
library(DT)
skincancer <- read_table("skincancer.txt")
datatable(skincancer,
           options = list(pageLength = 5, scrollY = "200px"))
```

Show  entries Search:

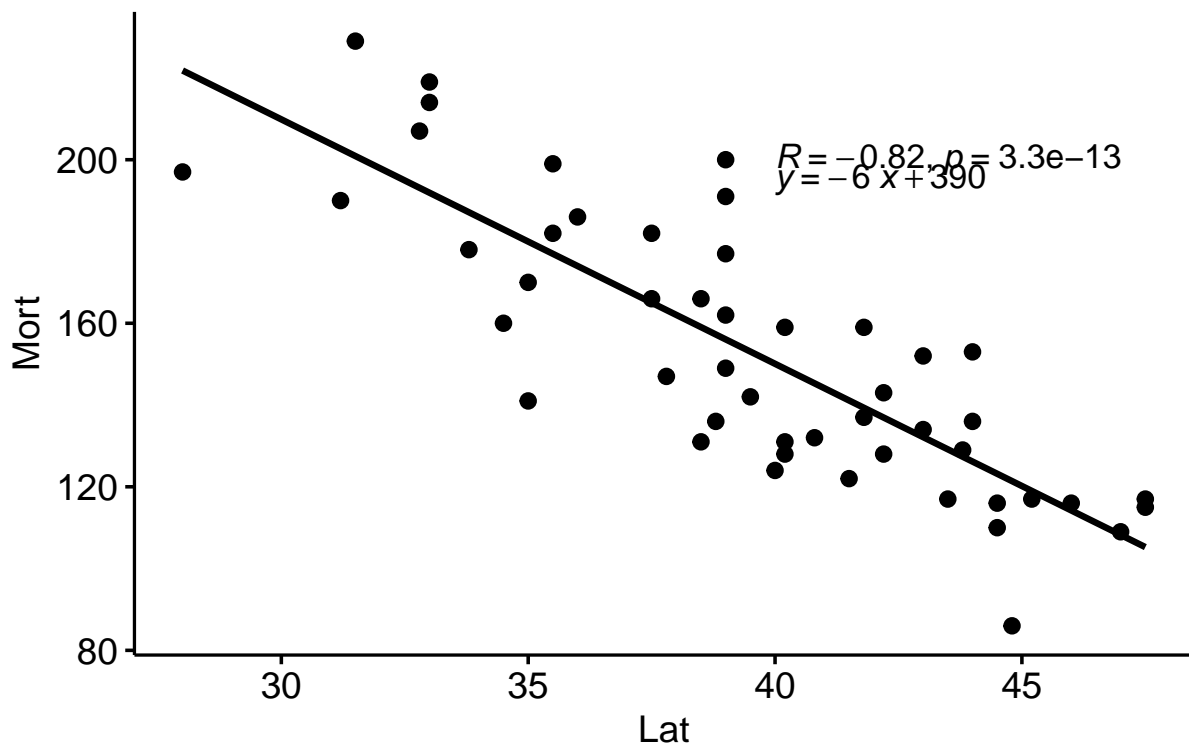
	State				
1	Alabama	33	219	1	87
2	Arizona	34.5	160	0	112
3	Arkansas	35	170	0	92.5
4	California	37.5	182	1	119.5
5	Colorado	39	149	0	105.5

Showing 1 to 5 of 49 entries

Previous  2 3 4 5 ... 10 Next

Note que viver nas latitudes mais altas do norte dos Estados Unidos, diminuiria a exposição aos raios nocivos do sol e, portanto, menos risco teria de morrer devido ao câncer de pele. O gráfico de dispersão suporta tal hipótese. Parece haver uma relação linear negativa entre latitude e mortalidade por câncer de pele, mas a relação não é perfeita. De fato, o enredo exibe alguma “tendência”, mas também exibe alguma “dispersão”. Portanto, é uma relação estatística, não determinística.





Alguns outros exemplos de relações estatísticas podem incluir:

- Altura e peso — à medida que a altura aumenta, você esperaria que o peso aumentasse, mas não perfeitamente.
- Álcool consumido e teor alcoólico no sangue — à medida que o consumo de álcool aumenta, você esperaria que o teor alcoólico no sangue aumentasse, mas não perfeitamente.
- Capacidade pulmonar vital e maços-ano de tabagismo — à medida que a quantidade de fumo aumenta (conforme quantificado pelo número de maços-ano de tabagismo), você esperaria que a função pulmonar (conforme quantificada pela capacidade pulmonar vital) diminuísse, mas não perfeitamente.
- Velocidade de direção e consumo de combustível — à medida que a velocidade de direção aumenta, você esperaria que o consumo de combustível diminuísse, mas não perfeitamente.

Portanto, vamos estudar as relações estatísticas entre uma variável de resposta  $y$  e uma variável preditora  $x$ !

### **3.1.2 Pressupostos do modelo**

### **3.1.3 Estimação dos parâmetros pelo método dos mínimos quadrados**

### **3.1.4 Propriedades dos estimadores**

### **3.1.5 Decomposição da Soma de Quadrados Total**

### **3.1.6 Tabela de ANOVA**

### **3.1.7 Coeficiente de Determinação**

### **3.1.8 Coeficiente de Determinação Ajustado para Graus de Liberdade**

### **3.1.9 Testes de Hipóteses sobre a inclinação e o intercepto**

### **3.1.10 Intervalos de Confiança para a inclinação e para o intercepto**

### **3.1.11 Intervalos de Confiança para a variância e para a média da variável resposta para um valor fixo da variável independente**

### **3.1.12 Intervalos de Previsão**

### **3.1.13 Teste para Falta de Ajustamento**

### **3.1.14 Análise de Resíduos**

### **3.1.15 Analisar dados usando o R**

## **3.2 Regressão Linear Múltipla**

Agora iremos admitir que  $X_1, X_2, \dots, X_k$  sejam variáveis independentes e  $Y$  a variável resposta. Dada uma amostra aleatória de  $n$  observações  $(x_{1i}, x_{2i}, \dots, x_{ki}, y_i)$ ,  $i = 1, 2, \dots, n$ , o modelo de regressão linear múltipla será dado por

$$E(Y_i | x_{1i}, x_{2i}, \dots, x_{ki}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}, \quad i = 1, 2, \dots, n,$$

ou

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \epsilon_i, \quad i = 1, 2, \dots, n,$$

em que  $n > k + 1$ .

Iremos considerar uma estrutura similar a do modelo de regressão linear simples. Especificamente, estamos considerando o seguinte:

- Modelo de regressão linear, ou **linear nos parâmetros**.
- Valores fixos de  $X$ .
- O termo de erro  $\epsilon_i$  tem valor médio zero.
- Homocedasticidade ou variância constante de  $\epsilon_i$ .
- Ausência de autocorrelação, ou de correlação serial, entre os termos de erro.
- Não há colinearidade exata entre as variáveis  $X$ .
- Ausência de viés de especificação.

### 3.2.1 Interpretação da equação de regressão múltipla

Considerando que temos apenas duas variáveis regressoras, então

$$E(Y_i|x_{1i}, x_{2i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}.$$

Desta forma, a equação acima fornece o **valor esperado ou média de  $Y$  condicional aos valores dados ou fixados de  $X_1$  e  $X_2$** .

Os coeficientes de regressão  $\beta_1$  e  $\beta_2$  são conhecidos como **coeficientes parciais de regressão** ou **coeficientes parciais angulares**. Seu significado é o seguinte:  $\beta_1$  mede a *variação* no valor médio de  $Y$ ,  $E(Y)$ , por unidade de variação em  $X_2$ , mantendo-se o valor de  $X_2$  constante. Em outras palavras, ele nos dá o efeito “direto” ou “liquido” de uma unidade de variação em  $X_2$  sobre o valor médio em  $Y$ , excluídos os efeitos que  $X_2$  possa ter sobre a média de  $Y$ . De modo análogo,  $\beta_2$  mede a variação do valor médio de  $Y$  por unidade de variação em  $X_2$ , mantendo-se constante o valor de  $X_1$ . Eles nos dá o efeito “direto” ou “liquido” de uma unidade de variação de  $X_2$  sobre o valor médio de  $Y$ , excluídos quaisquer efeitos que  $X_1$  possa ter sobre o valor médio de  $Y$ .

### 3.2.2 Abordagem Matricial

Por praticidade iremos utilizar a abordagem matricial, que no permitirar, entre outras coisas: i) encontrar o vetor de estimadores; ii) verificar as propriedades estatísticas dos estimadores; iii) obter a distribuição dos estimadores; qualquer que seja o número de variáveis independentes no modelo.

Sendo assim, podemos escrever o modelo de regressão linear múltipla como:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

que é conhecido como **modelo linear geral**.

Para determinarmos os estimadores de mínimos quadrados ordinários devemos minimizar

$$S(\beta) = \sum_{i=1}^n (\epsilon_i)^2 = \epsilon_1^2 + \dots + \epsilon_n^2 = \epsilon^\top \epsilon,$$

ou

$$S(\beta) = \epsilon^\top \epsilon = (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta).$$

### 3.2.3 Método dos Mínimos Quadrados Ordinários

Podemos abrir a expressão anterior da seguinte maneira

$$\begin{aligned} S(\beta) &= (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta) \\ &= (\mathbf{y}^\top - \beta^\top \mathbf{X}^\top)(\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\beta - \beta^\top \mathbf{X}^\top \mathbf{y} + \beta^\top \mathbf{X}^\top \mathbf{X}\beta. \end{aligned}$$

Observe que  $\mathbf{y}^\top \mathbf{X}\beta$  e  $\beta^\top \mathbf{X}^\top \mathbf{y}$  são escalares e

$$\mathbf{y}^\top \mathbf{X}\beta = (\beta^\top \mathbf{X}^\top \mathbf{y})^\top,$$

consequentemente  $\mathbf{y}^\top \mathbf{X}\beta = \beta^\top \mathbf{X}^\top \mathbf{y}$ . Desta forma,

$$S(\beta) = \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\beta + \beta^\top \mathbf{X}^\top \mathbf{X}\beta.$$

Nosso interesse, agora, é calcular  $\frac{\partial S(\beta)}{\partial \beta}$ . Temos que

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X}\beta.$$