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 determinística.

1. Circuferência = $\pi \times$ diâmetro.

2. **Lei de Hooke:** $Y = \alpha + \beta X$, em que Y é a quantidade de alogamento 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 constate, $P = \alpha/V$, em que P é a pressão, α é uma constante para cada gás e V é o volume do gás.

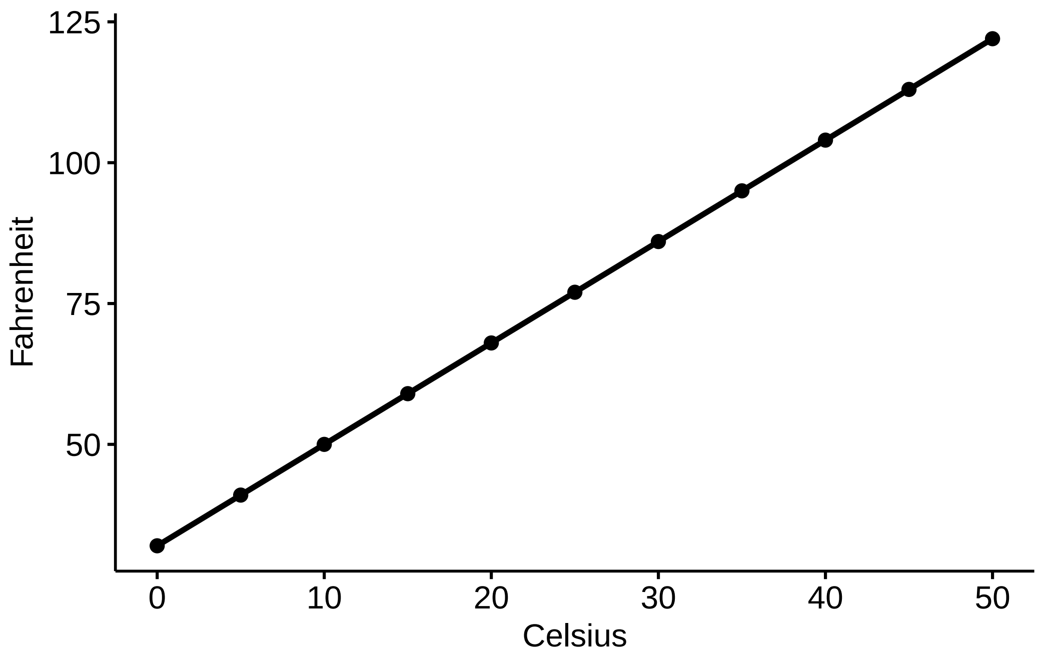
Para cada uma dessas relações determinísticas, a equação descreve exatamente a relação entre as duas variáveis. Esta disciplina não examina relacionamentos determinísticos. Em vez disso, estamos interessados em relações estatísticas, nas quais a relação entre as variáveis não é perfeita.

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 exeplo de relação estatística. A variável resposta Y é a mortalidade por cancêr 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"))
```

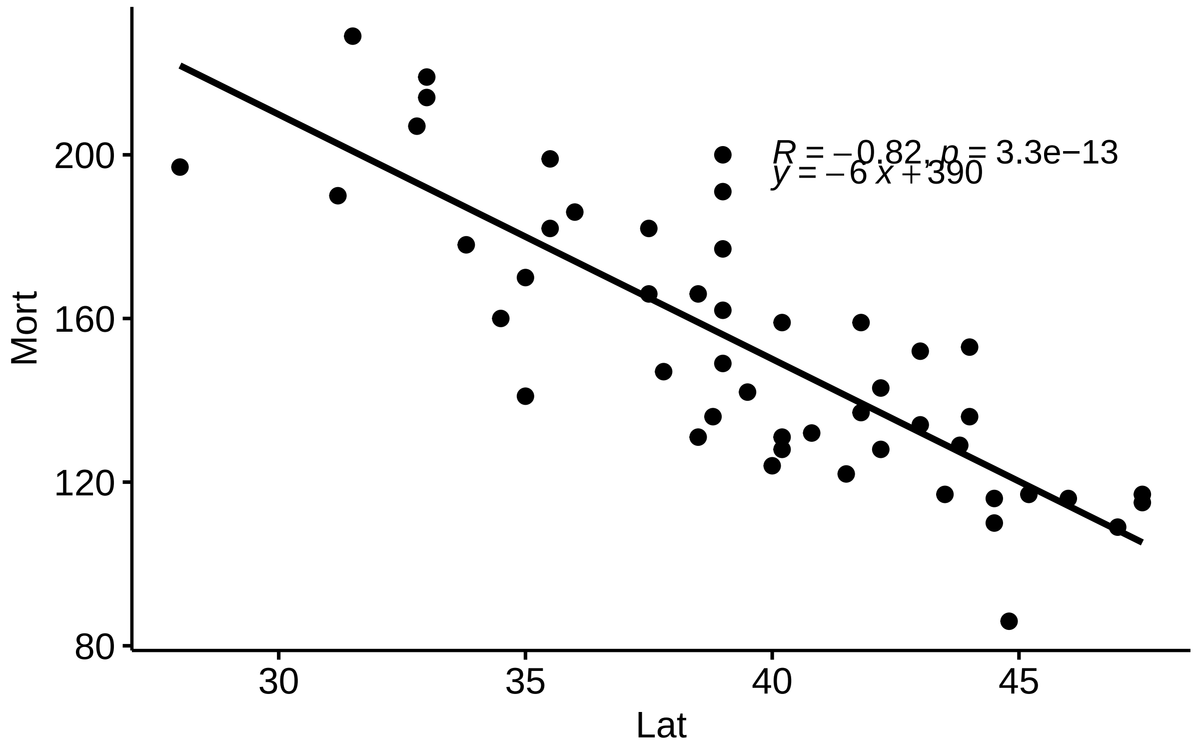
Show 5 entries Search:

	State				l
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 1 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

Observação:

Importante destacar que o termo regressão linear significa .blue[regressão linear nos parâmetros], ou seja, da forma

$$y_i = \alpha + \beta x_i^2 + u_i$$

ou da forma

$$\log(y_i) = \alpha + \beta \log(x_i) + u_i,$$

também são considerados **regressões lineares**.

O parâmetro

$$E(Y|X = x) = \alpha + \beta x$$

que representa a média da variável aleatória Y, condicionada a $X = x$, será estimada por

$$E(Y\widehat{|X} = x) = a + bx,$$

em que a e b são estimativas para α e β. A quantidade

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i), \quad i = 1, \dots, n,$$

é chamada de resíduo.

Assim, o valor e_i pode ser interpretado como o erro cometido por prever $y_i(i = 1, \dots, n)$ a partir de $\hat{y}_i$.

Voltando ao Exemplo (Semana 1)

Quais as estimativas do modelo de regressão linear simples de interesse?

$$\hat{y} = 390 - 6x.$$

$$y_i = E(Y|X = x_i) + u_i = \alpha + \beta x_i + u_i,$$

em que α é o intercepto e β é o coeficiente angular da reta de regressão.

Na prática, nem sempre α (intercepto) apresenta interpretação.

Como as estimativas devem ser interpretadas?

Voltando ao Exemplo (Semana 1)

$$\hat{y} = 390 - 6x.$$

- 390: valor médio de mortes por cancêr de pele em um estado com latitude central igual a zero. (Faz sentido essa interpretação?)
- -6: variação média no número de mortes quando aumenta-se a latitude em 1 unidade.

Exercício: Encontre a matriz hessiana e verifique sob quais condições a mesma é definida como positiva. Ainda, discuta se os estimadores encontrados geram o mínimo da função de interesse.

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

O modelo de regressão proposto está bem ajustado? Como medir a qualidade de ajuste do modelo?

> **Objetivo:** Construir uma medida que indique, mesmo que de modo imperfeito, a qualidade do ajuste do modelo de regressão.

- $y - \bar{y}$: erro ao se prever y pela média geral.
- $y - \hat{y}$: erro ao se prever y pelo valor estimado para $E(Y|X)$.
- $\hat{y} - \bar{y}$: "ganho" ao se prever y pelo valor estimado para $E(Y|X)$ em comparação ao se prever y pela média geral.
- Soma de Quadrados Total (SQT): $SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2$.
- Soma de Quadrados devido aos Resíduos (SQE): $SQE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.
- Soma de Quadrados devido ao modelo de regressão (SQReg): $SQReg = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$.

Resultado: $SQT = SQReg + SQE$

- Na SQT temos $n - 1$ graus de liberdade.
- Na SQE temos $n - 2$ graus de liberdade.