MODELOS LINEARES

Modelos Lineares Generalizados

Luan Fiorentin

12 de maio de 2021

Sumário

Introdução

Introdução

- **Inventário florestal** envolve a coleta de informações quali- quantitativas de variáveis da floresta.
- **Variáveis**:
  - Altura;
  - Diâmetro;
  - Presença de ataques de pragas;
  - Número de árvores por unidade amostral;
  - Entre outras.
- Fundamental reconhecer a **natureza** das variáveis.

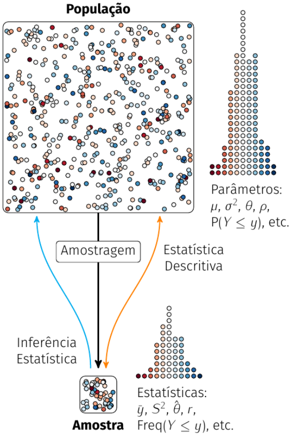

Figura 1. Inferência estatística. Extraído de Walmes Zeviani no Tikz.

Introdução

- Na modelagem estatística, procuramos estudar a **relação** entre duas ou mais variáveis Y e X:
$$Y = f(X).$$
- Cada observação é composta por duas partes:
  - **Previsível** (ou controlada).
  - **Aleatória** (não previsível ou não controlada).

$$\text{observação} = \text{previsível} + \text{aleatória}.$$

  - Parte previsível incorpora o **conhecimento sobre a variável**, sendo expressa por uma função matemática.
  - Parte aleatória deve obedecer algum **modelo de probabilidade**.

Introdução

- **Matemáticamente**, podemos descrever a relação como

$$y_i = \theta + \epsilon_i,$$

em que:

  - y_i é a observação i da variável resposta Y;
  - θ é um parâmetro de efeito fixo, comum a todas as observações;
  - ϵ_i é o erro da observação i, ou efeito resídual ou aleatório, sendo $\epsilon_i = y_i - \theta$.
  - ϵ_i pode ser resultante de outras variáveis que não foram controladas (ou não são controláveis). Logo, elas não estão explícitas no modelo.

Introdução - Exemplo 1

- **Exemplo**: Interesse em estimar a altura total de um povoamento florestal.
    - Foram mensuradas as variáveis altura, diâmetro e idade em nível de indivíduo.
    - ***Pergunta***: o que fazer com as árvores em que a altura não foi mensurada no inventário florestal?
    - ***Resposta***: estimar a altura das árvores.

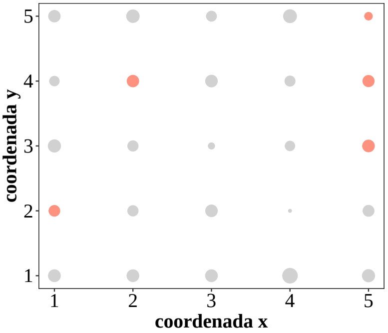

Figura 2. Árvores amostradas em colorido.

Introdução - Exemplo 1

- **Exemplo**: Amostra coletada:

altura	diâmetro
8.0	15.0
8.3	15.5
10.0	16.0
11.7	16.5
12.0	17.0

Podemos estimar a altura das árvores (y_i) de duas formas:

- Usando apenas os valores de y_i.
- Usando a relação de y_i com diâmetro das árvores (x_i).

- Modelo constante:

$$y_i = \theta + \epsilon_i$$

- Modelo linear simples:

$$y_i = \beta_0 + \beta_1 \text{diâmetro}_i + \epsilon_i$$

Introdução - Exemplo 1

- **Exemplo**: Considerando que a altura média das árvores é $\hat{\mu} = 10$ metros, então $\hat{y}_i$ pode ser descrita por:

$$\hat{y}_i = 10 + e_i,$$

em que:

  - $\hat{\theta} = \hat{\mu}$;
  - e_i determinará a altura de cada árvore em função de diversos fatores:

$$e_i = f(\text{diâmetro}, \text{sítio}, \text{idade}, \ldots).$$

Conforme relacionamos a altura com outras variáveis, ganhamos informação e **diminuímos o erro**.

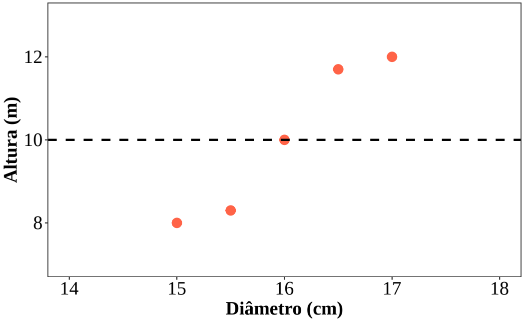

Figura 3. Modelagem da altura.

Introdução - Exemplo 1

- **Exemplo**: Considerando que a altura média das árvores é uma f(diâmetro), então $\hat{y}_i$ pode ser descrita por:

$$\hat{y}_i = -26.48 + 2.28\text{diâmetro}_i + e_i,$$

em que:

- $\hat{\theta} = \hat{\mu}$;
- e_i determinará a altura de cada árvore em função de diversos fatores:

$$e_i = f(\text{sítio}, \text{idade}, \ldots).$$

Conforme relacionamos a altura com outras variáveis, ganhamos informação e **diminuímos o erro**.

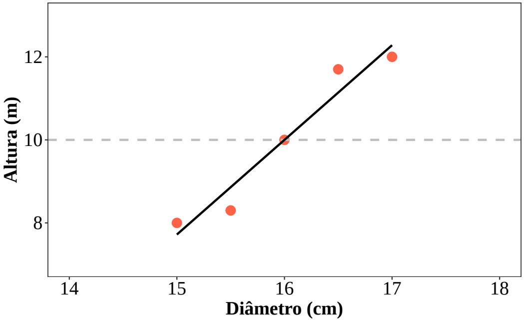

Figura 4. Modelagem da altura e o diâmetro.

Modelos Lineares

Modelos Lineares

- É uma classe de modelos onde o **parâmetro** ocorre de forma **linear**.
- Casos **especiais**:
  - Análise de variância (ANOVA).
  - Análise de regressão linear.
- **Análise de variância**: compara simultâneamente k médias.
- **Análise de regressão linear**: estuda a relação entre variáveis Y e X:
  - Regressão linear simples.
  - Regressão linear múltipla.

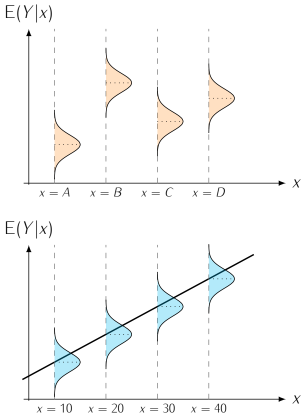

Figura 5. Representações esquemáticas dos modelos de análise de variância e regressão linear simples. Extraído de DEST.

Regressão Linear Simples

Regressão Linear Simples

- **Modelo de regressão linear simples** é definido por uma reta que estabelece a relação entre uma variável resposta (y_i) e uma explicativa (x_i):

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

  em que:
  - i é o índice da observação;
  - β_0 é o intercepto da reta;
  - β_1 é o coeficiente angular da reta;
  - ϵ é o erro aleatório.
- Definimos que os erros seguem uma **distribuição normal** de média zero e variância constante, ou seja, $\epsilon \sim N(0, \sigma^2)$.
- Ainda, supomos que erros associados a diferentes observações sejam **não correlacionados**, então $Cov(\epsilon_i, \epsilon_i^{'}) = 0$.

Regressão Linear Simples

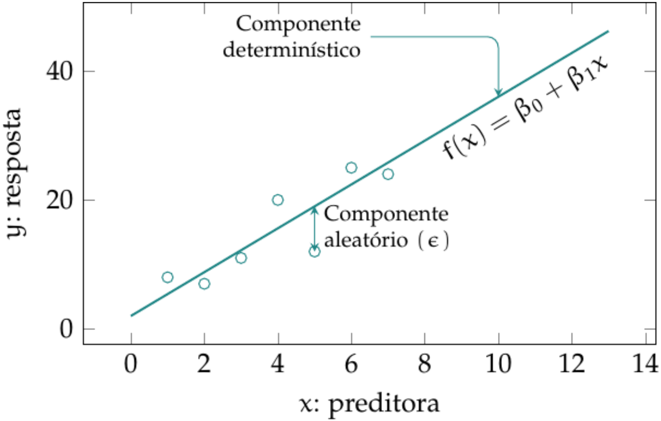

Figura 6. Componentes do modelo. Extraído de Walmes Zeviani no Tikz.

Regressão Linear Simples

- **Média** de y_i:

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i.$$

- **Variância** de y_i:

$$Var(y_i|x_i) = \sigma^2.$$

- **Interpretação** dos parâmetros:
  - β_1 expressa a alteração no valor esperado de y_i associado ao acréscimo de uma unidade em x_i.
  - β_0 é o valor esperado de y_i quando $x_i = 0$.

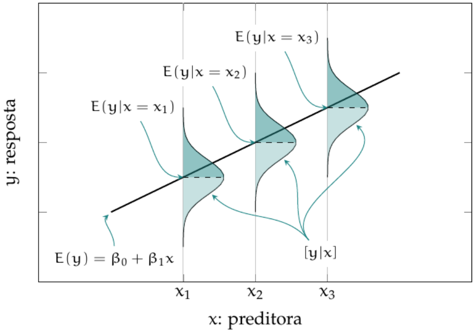

Figura 7. Valor esperado da resposta. Extraído de Walmes Zeviani no Tiz.

Regressão Linear Simples

- **Valores preditos**: A predição de valores estimados para a resposta depende de valores da variável preditora:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

- **Resíduos** são erros de predição ao se utilizar um modelo para descrever a relação entre variáveis:

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

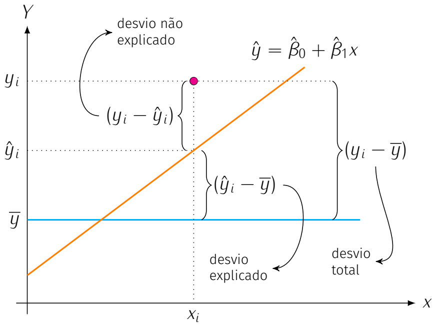

Figura 8. Resíduos do modelo de regressão linear simples. Extraído de Walmes Zeviani no Tikz.

Regressão Linear Simples

- **Estimação dos parâmetros**: Note que o resíduo (ϵ_i) é função dos parâmetros desconhecidos β_0 e β_1.
- Precisamos de um **critério** para definir valores ótimos de parâmetros.
- Critério de **mínimos quadrados** consiste em minimizar a soma de quadrados dos resíduos (SSE):

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

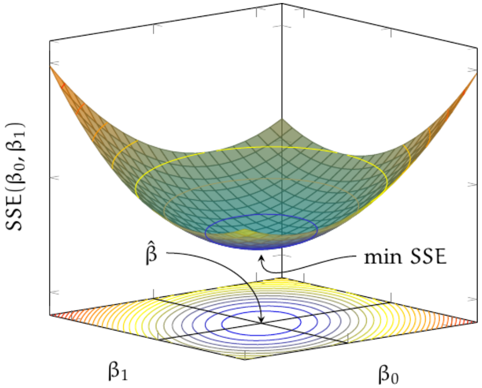

Figura 9. Minimização da SSE. Extraído de Walmes Zeviani no Tikz

Regressão Linear Simples

- Abordagem padrão é usar **cálculo diferencial**:
  - Obter **vetor gradiente** (vetor de derivadas parciais) de β_0 e β_1;
  - Resolver o **sistema de equações lineares** (igualar a zero e isolar β_0 e β_1):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2},$$

sendo $\bar{x}$ e $\bar{y}$ a média de cada variável. Já a variância é

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}.$$

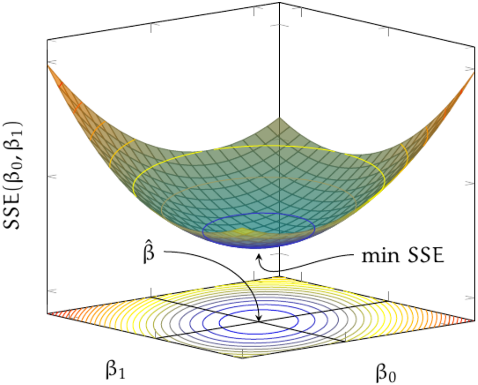

Figura 10. Minimização da SSE. Extraído de Walmes Zeviani no Tikz

Regressão Linear Simples - Exemplo 2

- O problema consiste em **estimar a altura total** de árvores de um inventário florestal realizado em determinada idade.
- **Dados** disponíveis:
  - Altura (H) $\rightarrow$ Resposta.
  - Diâmetro (D) $\rightarrow$ Preditora.
- Modelo **teórico**:

$$H_i = \beta_0 + \beta_1 D_i + \epsilon_i.$$

- **Valor esperado**:

$$E(H_i) = \hat{\beta}_0 + \hat{\beta}_1 D_i.$$

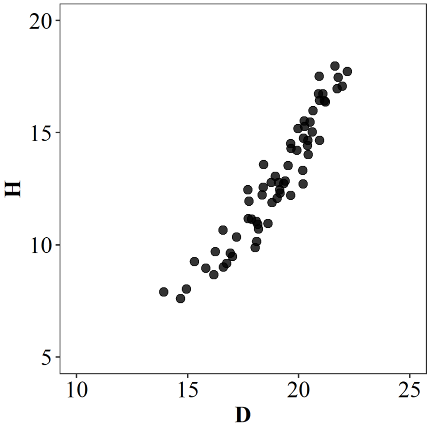

Figura 11. Relação entre altura (H) e diâmetro (D).

Regressão Linear Simples - Exemplo 2

- O problema consiste em **estimar a altura total** de árvores de um inventário florestal realizado em determinada idade.
- **Dados** disponíveis:
  - Altura (H) $\rightarrow$ Resposta.
  - Diâmetro (D) $\rightarrow$ Preditora.
- Modelo **teórico**:

$$H_i = \beta_0 + \beta_1 D_i + \epsilon_i.$$

- **Valor esperado**:

$$E(H_i) = \hat{\beta}_0 + \hat{\beta}_1 D_i.$$

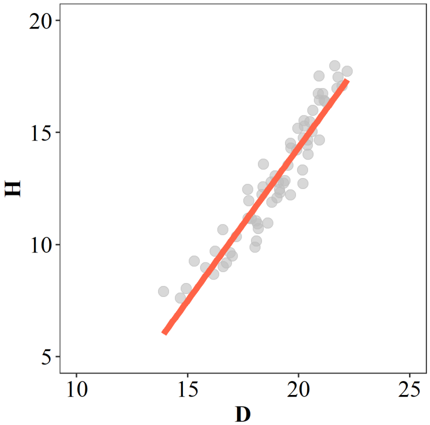

Figura 12. Modelagem da relação entre altura (H) e diâmetro (D).

Regressão Linear Múltipla

- **Modelo de regressão linear múltiplo** é definido por função que estabelece a relação entre uma variável resposta (y_i) e múltiplas explicativas ($x_{i1}, \ldots, x_{ip}$):

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i,$$

  em que:
  - i é o índice da observação;
  - β_0 é o intercepto;
  - β_j é o parâmetro, sendo $j = 1, \ldots, p$;
  - x_j é a variável preditora, sendo $j = 1, \ldots, p$;
  - ϵ é o erro aleatório.
- Definimos que os erros seguem uma **distribuição normal** de média zero e variância constante, ou seja, $\epsilon \sim N(0, \sigma^2)$.
- Ainda, supomos que erros associados a diferentes observações sejam **não correlacionados**, então $Cov(\epsilon_i, \epsilon_i^{'}) = 0$.

Regressão Linear Múltipla

- **Média** de y_i:

$$E(y_i|\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})^{'})) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

- **Variância** de y_i:

$$Var(y_i|\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})^{'})) = \sigma^2.$$

- **Interpretação** dos parâmetros:
  - β_j representa a alteração esperada na resposta (y_i) para uma unidade a mais em x_{ij}, quando as demais preditoras são mantidas fixas.

Regressão Linear Múltipla

- **Valores preditos**: A predição de valores estimados para a resposta depende de valores das variáveis preditoras:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}.$$

- **Resíduos** são erros de predição ao se utilizar um modelo para descrever a relação entre variáveis:

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

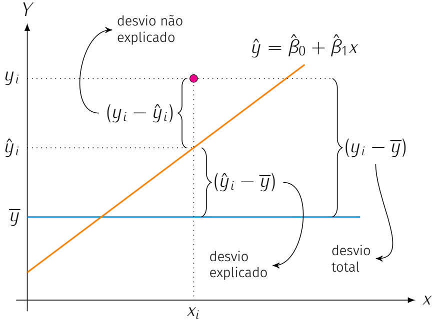

Figura 13. Resíduos do modelo de regressão linear múltipla. Extraído de Walmes Zeviani no Tikz.

Regressão Linear Múltipla

- **Notação matricial**: Modelo de regressão linear múltiplo é dado por

$$\boldsymbol{y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

em que:

  - $\boldsymbol{y} = (y_1, \ldots, y_n)'$ de dimensão $n \times 1$;
  - $\boldsymbol{x} = (1_i, x_{i1}, \ldots, x_{np})'$ de dimensão $n \times (p + 1)$;
  - $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ de dimensão $(p + 1) \times 1$;
  - $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ de dimensão $n \times 1$.

Regressão Linear Múltipla

- **Estimação dos parâmetros**: Estimação de mínimos quadrados baseia-se, novamente, na determinação de $\beta_0, \beta_1, \ldots, \beta_p$ que minimizem a soma de quadrados de resíduos dada por
$$SSE(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}))^2.$$
Em notação matricial: $SSE(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})^{'}(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})$.
- Estimador de mínimos quadrados em notação matricial é dado por
$$\hat{\boldsymbol{\beta}} = (\boldsymbol{x}^{'}\boldsymbol{x})^{-1}\boldsymbol{x}^{'}\boldsymbol{y}.$$
- Estimador de variância é dado por
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p}.$$

Regressão Linear Múltipla - Exemplo 3

- O problema consiste em **estimar o volume total** de árvores de um inventário florestal realizado em determinada idade.
- **Dados** disponíveis:
  - Volume (H) $\rightarrow$ Resposta.
  - Diâmetro (D) $\rightarrow$ Preditora.
- Modelo **teórico**:

$$V_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \epsilon_i.$$

- **Valor esperado**:

$$E(H_i) = \hat{\beta}_0 + \hat{\beta}_1 D_i + \hat{\beta}_2 D_i^2.$$

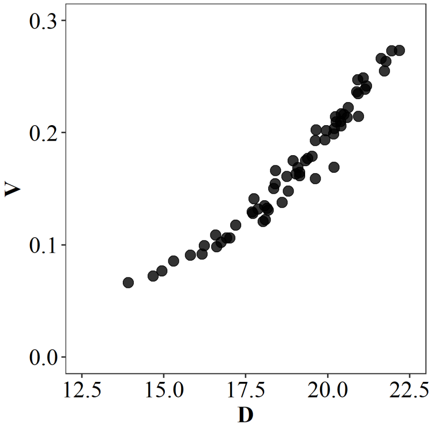

Figura 14. Relação entre volume (V) e diâmetro (D).

Regressão Linear Múltipla - Exemplo 3

- O problema consiste em **estimar o volume total** de árvores de um inventário florestal realizado em determinada idade.
- **Dados** disponíveis:
  - Volume (H) $\rightarrow$ Resposta.
  - Diâmetro (D) $\rightarrow$ Preditora.
- Modelo **teórico**:

$$V_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2 + \epsilon_i.$$

- **Valor esperado**:

$$E(H_i) = \hat{\beta}_0 + \hat{\beta}_1 D_i + \hat{\beta}_2 D_i^2.$$

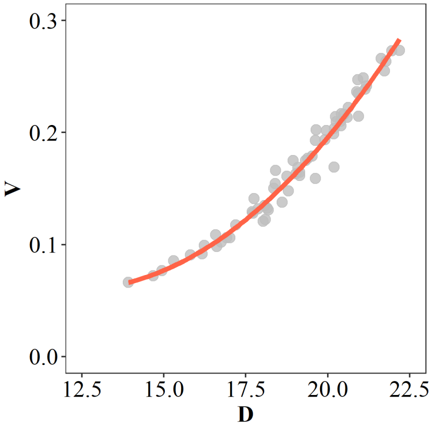

Figura 15. Modelagem da relação entre volume (V) e diâmetro (D).

Considerações finais

Considerações finais

- **Modelos de regressão** são fundamentais para solução de problemas florestais.
- **Regressão linear simples** é método de análise bastante intutitivo.
- **Regressão linear múltipla** é uma generalização dos modelos lineares.