# Modelo de Regressão Linear Simples

Prof. Juvêncio Santos Nobre

Departamento de Estatística e Matemática Aplicada

Universidade Federal do Ceará-Brasil

http://www.dema.ufc.br/∼juvencio

DEMA-UFC

Capital do **Ceará**, agosto de 2022

# Conteúdo

# MRLS

- O modelo de regressão linear simples (MRLS) admite a seguinte forma funcional

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \ldots, n, \tag{1}$$

em que:

  - $y_i$ ($x_i$) denota o valor da variável resposta (explicativa) referente ao $i$-ésimo elemento da amostra.
  - $\beta_0$ e $\beta_1$ são parâmetros desconhecidos, denominados parâmetros (coeficientes) de regressão.
  - $e_i$ representa a fonte de variação associada ao $i$-ésimo elemento da amostra.

# MRLS

- O modelo de regressão linear simples (MRLS) admite a seguinte forma funcional

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \ldots, n, \tag{1}$$

em que:

  - $y_i$ ($x_i$) denota o valor da variável resposta (explicativa) referente ao $i$-ésimo elemento da amostra.
  - $\beta_0$ e $\beta_1$ são parâmetros desconhecidos, denominados parâmetros (coeficientes) de regressão.
  - $e_i$ representa a fonte de variação associada ao $i$-ésimo elemento da amostra.

# MRLS

- O modelo de regressão linear simples (MRLS) admite a seguinte forma funcional

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \ldots, n, \tag{1}$$

em que:

- $y_i$ ($x_i$) denota o valor da variável resposta (explicativa) referente ao $i$-ésimo elemento da amostra.
- $\beta_0$ e $\beta_1$ são parâmetros desconhecidos, denominados parâmetros (coeficientes) de regressão.
- $e_i$ representa a fonte de variação associada ao $i$-ésimo elemento da amostra.

# MRLS

- O modelo de regressão linear simples (MRLS) admite a seguinte forma funcional

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \ldots, n, \tag{1}$$

em que:

  - $y_i$ ($x_i$) denota o valor da variável resposta (explicativa) referente ao $i$-ésimo elemento da amostra.
  - $\beta_0$ e $\beta_1$ são parâmetros desconhecidos, denominados parâmetros (coeficientes) de regressão.
  - $e_i$ representa a fonte de variação associada ao $i$-ésimo elemento da amostra.

# MRLS

- Ao estabelecer o MRLS, pressupomos que:

  i) A função de regressão é linear (nos parâmetros). É comum, apesar de formalmente incorreta, nos textos aparecer a relação entre $y_i$ e $x_i$ é linear nos parâmetros.

  ii) Os valores de $x_i$ são fixos, i.e., $x_i$ não é uma variável aleatória.

  iii) $\mathbb{E}[e_i] = 0$, $\forall i = 1, \ldots, n$. Na verdade, tal suposição deveria ser escrita como (o que acaba implicando a anterior) $\mathbb{E}[e_i|x_i] = 0, \forall i = 1, \ldots, n$.

  iv) Para um dado valor de $x_i$, a variância da fonte de variação é constante, i.e.,

  $$\mathrm{Var}[e_i] = \mathbb{E}[e_i^2] = \sigma^2, \forall i = 1, \ldots, n \quad (\textbf{Homoscesdaticidade}).$$

  Na verdade, tal suposição deveria ser escrita como

  $$\mathrm{Var}[y_i|x_i] = \mathrm{Var}[e_i|x_i] = \sigma^2, \forall i = 1, \ldots, n.$$

  v) A fonte de variação associada a uma observação é não-correlacionada com a fonte de variação associada de outra observação, i.e.,

  $$\mathrm{Cov}(e_i, e_j) = \mathbb{E}[e_i e_j] = 0, \forall i \neq j.$$

# MRLS

- As suposições iv) e v) podem ser reescritas de sucintamente da seguinte forma

$$\mathrm{Cov}(e_i, e_j) = \sigma^2 \mathbb{1}(i = j), \forall i, j = 1 \ldots, n.$$

- Perceba que no MRLS (1) assume-se essencialmente que a fonte de variação está relacionada somente a variável resposta, i.e, a variável explicativa é medida sem erro, ou seja, com completa exatidão. Isso é razoável no contexto prático?
- Se tivermos uma fonte de variação também associada a variável explicativa $x_i$, teremos essencialmente um *modelo com erro de medida/erro nas variáveis*.

# MRLS

- As suposições iv) e v) podem ser reescritas de sucintamente da seguinte forma

$$\text{Cov}(e_i, e_j) = \sigma^2 \mathbb{1}(i = j), \forall i, j = 1 \ldots, n.$$

- Perceba que no MRLS (1) assume-se essencialmente que a fonte de variação está relacionada somente a variável resposta, i.e, a variável explicativa é medida sem erro, ou seja, com completa exatidão. Isso é razoável no contexto prático? 😵

- Se tivermos uma fonte de variação também associada a variável explicativa $x_i$, teremos essencialmente um *modelo com erro de medida/erro nas variáveis*. 😎

# MRLS

- As suposições iv) e v) podem ser reescritas de sucintamente da seguinte forma

$$\mathrm{Cov}(e_i, e_j) = \sigma^2 \mathbb{1}(i = j), \forall i, j = 1 \ldots, n.$$

- Perceba que no MRLS (1) assume-se essencialmente que a fonte de variação está relacionada somente a variável resposta, i.e, a variável explicativa é medida sem erro, ou seja, com completa exatidão. Isso é razoável no contexto prático?
- Se tivermos uma fonte de variação também associada a variável explicativa $x_i$, teremos essencialmente um *modelo com erro de medida/erro nas variáveis*.

# MRLS

- Para efeito de **inferência de segunda ordem** exata, i.e., construção de IC, testes de hipóteses, é comum considerar também que

$$e_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n.$$

- Lembrando, que **correlação nula** implica independência sob a suposição de normalidade multivariada, então usando as suposições iv) e v) adicionada com a suposição acima, temos

$$e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, n.$$

- Usando o fato que a distribuição normal é **fechada** por transformações lineares, então sob as suposições usuais do MRLS adicionada a suposição de normalidade, tem-se

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), i = 1, \ldots, n.$$

# MRLS

- Para efeito de **inferência de segunda ordem** exata, i.e., construção de IC, testes de hipóteses, é comum considerar também que

$$e_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n.$$

- Lembrando, que **correlação nula** implica independência sob a suposição de normalidade multivariada, então usando as suposições iv) e v) adicionada com a suposição acima, temos

$$e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, n.$$

- Usando o fato que a distribuição normal é **fechada** por transformações lineares, então sob as suposições usuais do MRLS adicionada a suposição de normalidade, tem-se

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), i = 1, \ldots, n.$$

# MRLS

- Para efeito de **inferência de segunda ordem** exata, i.e., construção de IC, testes de hipóteses, é comum considerar também que

$$e_i \sim \mathcal{N}(0, \sigma^2), i = 1, \ldots, n.$$

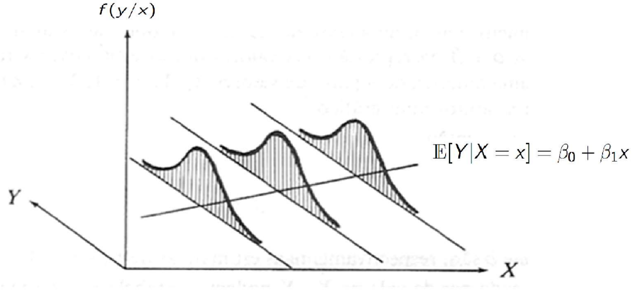
- Lembrando, que **correlação nula** implica independência sob a suposição de normalidade multivariada, então usando as suposições iv) e v) adicionada com a suposição acima, temos

$$e_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, n.$$

- Usando o fato que a distribuição normal é **fechada** por transformações lineares, então sob as suposições usuais do MRLS adicionada a suposição de normalidade, tem-se

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), i = 1, \ldots, n.$$

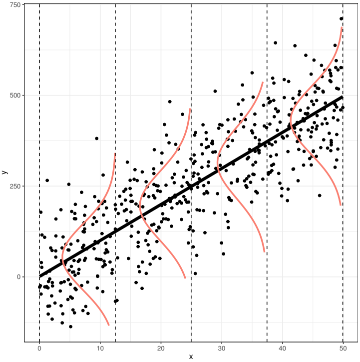
# MRLS - Ilustração gráfica




Fonte: Hoffman (2006, Análise de regressão)

# MRLS - Ilustração gráfica

Figura: Ilustração gráfica para um exemplo de dados simulados usando o ggplot2.



# MRLS - Interpretação dos parâmetros

- Sob as suposições usuais do MRLS, tem-se

$$\mathbb{E}[y_i|X = x] = \beta_0 + \beta_1 x, i = 1, \ldots, n.$$

Logo:

- $\beta_0 = \mathbb{E}[y_i|X = 0]$.
- É válido ressaltar que quando a amplitude amostral não inclui o zero (ou quando não fizer sentido considerar $x = 0$) , então $\beta_0$ não possui interpretação prática, sendo necessário centralizar a variável explicativa para tal.
- $\beta_1 = \mathbb{E}[y_i|X = a + 1] - \mathbb{E}[y_i|X = a], \forall a \in \mathbb{R}$, i.e., $\beta_1$ representa a variação no valor esperado da variável resposta, quando a variável explicativa é acrescida de uma unidade de medida.

# MRLS - Interpretação dos parâmetros

- Sob as suposições usuais do MRLS, tem-se

$$\mathbb{E}[y_i|X = x] = \beta_0 + \beta_1 x, i = 1, \ldots, n.$$

Logo:

- $\beta_0 = \mathbb{E}[y_i|X = 0]$.
- É válido ressaltar que quando a amplitude amostral não inclui o zero (ou quando não fizer sentido considerar $x = 0$) , então $\beta_0$ não possui interpretação prática, sendo necessário centralizar a variável explicativa para tal.
- $\beta_1 = \mathbb{E}[y_i|X = a + 1] - \mathbb{E}[y_i|X = a], \forall a \in \mathbb{R}$, i.e., $\beta_1$ representa a variação no valor esperado da variável resposta, quando a variável explicativa é acrescida de uma unidade de medida.

# MRLS - Interpretação dos parâmetros

- Sob as suposições usuais do MRLS, tem-se

$$\mathbb{E}[y_i|X = x] = \beta_0 + \beta_1 x, i = 1, \ldots, n.$$

Logo:

- $\beta_0 = \mathbb{E}[y_i|X = 0]$.
- É válido ressaltar que quando a amplitude amostral não inclui o zero (ou quando não fizer sentido considerar $x = 0$) , então $\beta_0$ não possui interpretação prática, sendo necessário centralizar a variável explicativa para tal.
- $\beta_1 = \mathbb{E}[y_i|X = a + 1] - \mathbb{E}[y_i|X = a], \forall a \in \mathbb{R}$, i.e., $\beta_1$ representa a variação no valor esperado da variável resposta, quando a variável explicativa é acrescida de uma unidade de medida.

# MRLS - Interpretação dos parâmetros

- Sob as suposições usuais do MRLS, tem-se

$$\mathbb{E}[y_i|X = x] = \beta_0 + \beta_1 x, i = 1, \ldots, n.$$

Logo:

- $\beta_0 = \mathbb{E}[y_i|X = 0]$.
- É válido ressaltar que quando a amplitude amostral não inclui o zero (ou quando não fizer sentido considerar $x = 0$) , então $\beta_0$ não possui interpretação prática, sendo necessário centralizar a variável explicativa para tal.
- $\beta_1 = \mathbb{E}[y_i|X = a + 1] - \mathbb{E}[y_i|X = a], \forall a \in \mathbb{R}$, i.e., $\beta_1$ representa a variação no valor esperado da variável resposta, quando a variável explicativa é acrescida de uma unidade de medida.

# Exemplos - Interpretação dos parâmetros

**Exemplo 1:** Para os casos abaixo, apresente interpretações práticas dos parâmetros do MRLS:

i) Renda vs. anos estudados (efetivos).

ii) Peso vs altura.

iii) Tempo de processamento vs # de faturas.

iv) Faturamento da empresa vs investimento com propaganda.

v) Pressão arterial sistólica (mmHg) vs idade (anos).