

# Variável Binária e Interação

Gilberto A. Paula

Departamento de Estatística  
IME-USP, Brasil  
giapaula@ime.usp.br

2<sup>o</sup> Semestre 2023

- 1 Introdução
- 2 Variável Explicativa Binária
- 3 Variável Explicativa Categórica
- 4 Referências

## Objetivos

Neste material serão apresentados os seguintes conceitos:

- Variável Explicativa Binária
- Ausência e Presença de Interação
- Variável Explicativa Categórica
- Ausência e Presença de Interação

- 1 Introdução
- 2 Variável Explicativa Binária**
- 3 Variável Explicativa Categórica
- 4 Referências

## Ausência de Interação

Supor o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

em que  $y_1, \dots, y_n$  são valores observados da variável resposta,  $x_{i2}$  representa os valores de uma variável aleatória binária tal que

$$x_{i2} = \begin{cases} 1 & \text{grupo A} \\ 0 & \text{grupo B,} \end{cases}$$

enquanto  $x_{i3}$  representa valores observados de uma variável contínua e  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ , para  $i = 1, \dots, n$ .

## Ausência de Interação

Portanto, tem-se dois submodelos de regressão

- (Grupo A)  $y_i = \beta_1 + \beta_2 + \beta_3 x_{i3} + \epsilon_i$
- (Grupo B)  $y_i = \beta_1 + \beta_3 x_{i3} + \epsilon_i$

com valores esperados

- $E_A(Y_i | x_{i3}) = \beta_1 + \beta_2 + \beta_3 x_{i3}$
- $E_B(Y_i | x_{i3}) = \beta_1 + \beta_3 x_{i3},$

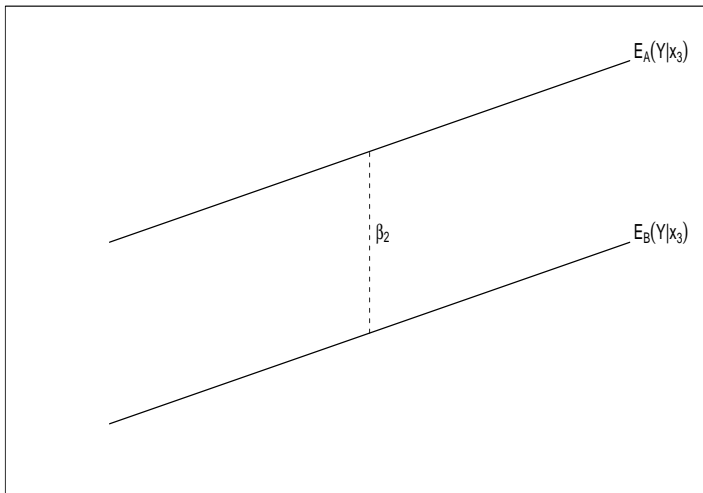
para  $i = 1, \dots, n.$

Assim,  $E_A(Y_i | x_{i3}) - E_B(Y_i | x_{i3}) = \beta_2$ , ausência de interação entre as variáveis explicativas  $X_2$  e  $X_3$ .

## Definição

A diferença entre os valores esperados entre dois níveis (valores) quaisquer de um fator (variável) **não muda** à medida que variam os níveis do outro fator (outra variável). Neste modelo tem-se que  $\beta_2$  **não muda à medida que variam os valores de  $X_3$** .

# Ilustração Ausência de Interação





## Ausência de Interação

Supondo que o grupo A tem  $n_1$  elementos e o grupo B  $n_2$  elementos o modelo com ausência de interação entre a variável binária  $X_2$  e a variável contínua  $X_3$  pode ser expresso na forma alternativa:

$$y_{ij} = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{ij3} + \epsilon_{ij},$$

em que  $j = 1, \dots, n_j$  e  $i = 1, 2$ .

## Ausência de Interação

Em forma matricial o modelo fica dado por

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

em que  $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top)^\top$  com  $\mathbf{y}_1 = (y_{11}, \dots, y_{1n_1})^\top$ ,  $\mathbf{y}_2 = (y_{21}, \dots, y_{2n_2})^\top$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$  e matriz  $\mathbf{X}$  de dimensão  $(n_1 + n_2) \times 3$  dada por

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & x_{113} \\ \vdots & \vdots & \vdots \\ 1 & 1 & x_{1n_13} \\ 1 & 0 & x_{213} \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{2n_23} \end{bmatrix}.$$

## Presença de Interação

Supor agora o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2} x_{i3} + \epsilon_i,$$

em que  $y_1, \dots, y_n$  são valores observados da variável resposta,  $x_{i2}$  representa os valores de uma variável aleatória binária tal que

$$x_{i2} = \begin{cases} 1 & \text{grupo A} \\ 0 & \text{grupo B,} \end{cases}$$

enquanto  $x_{i3}$  representa valores observados de uma variável contínua e  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ , para  $i = 1, \dots, n$ .

## Presença de Interação

Portanto, tem-se dois submodelos de regressão

- (Grupo A)  $y_i = \beta_1 + \beta_2 + \beta_3 X_{i3} + \beta_4 X_{i3} + \epsilon_i$
- (Grupo B)  $y_i = \beta_1 + \beta_3 X_{i3} + \epsilon_i$

com valores esperados

- $E_A(Y_i|X_{i3}) = \beta_1 + \beta_2 + \beta_3 X_{i3} + \beta_4 X_{i3}$
- $E_B(Y_i|X_{i3}) = \beta_1 + \beta_3 X_{i3},$

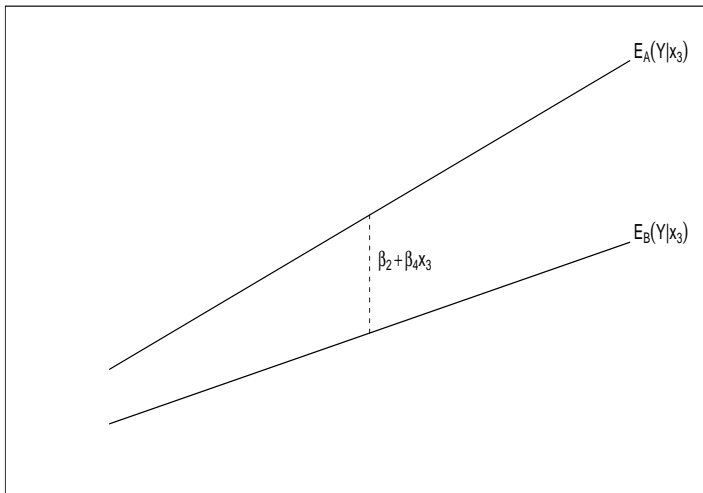
para  $i = 1, \dots, n.$

Assim,  $E_A(Y_i|X_{i3}) - E_B(Y_i|X_{i3}) = \beta_2 + \beta_4 X_{i3}$ , presença de interação entre as variáveis explicativas  $X_2$  e  $X_3$ .

## Definição

A diferença entre os valores esperados entre dois níveis (valores) quaisquer de um fator (variável) **não é constante** à medida que variam os níveis do outro fator (outra variável). Neste modelo tem-se que **a diferença entre os valores esperados para os grupos A e B depende dos valores da variável  $X_3$ .**

# Ilustração Presença de Interação



## Presença de Interação

Supondo que o grupo A tem  $n_1$  elementos e o grupo B  $n_2$  elementos o modelo com presença de interação entre a variável binária  $X_2$  e a variável contínua  $X_3$  pode ser expresso na forma alternativa:

$$y_{ij} = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{ij3} + \beta_4 X_{i2} X_{ij3} + \epsilon_{ij},$$

em que  $j = 1, \dots, n_j$  e  $i = 1, 2$ .

## Presença de Interação

Em forma matricial o modelo fica dado por

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

em que  $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top)^\top$  com  $\mathbf{y}_1 = (y_{11}, \dots, y_{1n_1})^\top$ ,  $\mathbf{y}_2 = (y_{21}, \dots, y_{2n_2})^\top$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^\top$  e matriz  $\mathbf{X}$  de dimensão  $(n_1 + n_2) \times 4$  dada por

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & x_{113} & x_{113} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & x_{1n_13} & x_{1n_13} \\ 1 & 0 & x_{213} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{2n_23} & 0 \end{bmatrix}.$$



- 1 Introdução
- 2 Variável Explicativa Binária
- 3 Variável Explicativa Categórica**
- 4 Referências

## Formulação do Modelo

Supor variável explicativa com três níveis

$$X = \begin{cases} 1 & \text{grupo A} \\ 2 & \text{grupo B} \\ 3 & \text{grupo C.} \end{cases}$$

Um maneira de representar essa variável explicativa num modelo de regressão é atribuindo a cada grupo uma variável binária:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i,$$

em que  $y_1, \dots, y_n$  denotam os valores observados da variável resposta,  $x_{i1}$ ,  $x_{i2}$  e  $x_{i3}$  são os valores observados das variáveis binárias representando os grupos e  $\epsilon_j \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ , para  $j = 1, \dots, n$ .

## Formulação do Modelo

Supondo que os grupos A, B e C têm  $n_1$ ,  $n_2$  e  $n_3$  elementos, respectivamente, o modelo pode ser expresso na forma matricial

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

em que  $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \mathbf{y}_3^\top)^\top$  com  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^\top$ , para  $i = 1, 2, 3$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$  e matriz  $\mathbf{X}$  de dimensão  $(n_1 + n_2 + n_3) \times 4$ .

## Formulação do Modelo

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} .$$

Note que a matriz  $\mathbf{X}$  **não tem posto coluna completo**, a 1<sup>a</sup> coluna é a soma das outras três colunas.

## Formulação do Modelo

Uma solução é reduzir o número de colunas da matriz modelo impondo alguma restrição nos parâmetros. Procedimentos mais utilizados:

- Restrição nos Parâmetros:  $\beta_1 + \beta_2 + \beta_3 = 0$ , que implica em  $\beta_1 = -\beta_2 - \beta_3$ .
- Casela de Referência: um dos coeficientes é fixado como sendo zero. Por exemplo, fazendo  $\beta_1 = 0$  o grupo A será denominado casela de referência.

Nesses dois casos  $\beta = (\beta_0, \beta_2, \beta_3)^\top$  e a matriz modelo terá dimensão  $n \times 3$  com posto coluna completo.

## Formulação do Modelo

Matriz modelo quando  $\beta_1 = -\beta_2 - \beta_3$ :

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & -1 & -1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} .$$

## Formulação do Modelo

Matriz modelo quando  $\beta_1 = 0$ :

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} .$$

## Ausência de Interação

O modelo com casela de referência **no grupo A** pode ser expresso na seguinte forma:

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i,$$

em que  $y_1, \dots, y_n$  denotam os valores observados da variável resposta,  $x_{i2}$  e  $x_{i3}$  são valores de variáveis binárias representando os grupos B e C, respectivamente, enquanto  $x_{i4}$  representa os valores observados de uma variável contínua e  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ , para  $i = 1, \dots, n$ .



## Ausência de Interação

Portanto, tem-se três submodelos

- (Grupo A)  $y_i = \beta_0 + \beta_4 x_{i4} + \epsilon_i$
- (Grupo B)  $y_i = \beta_0 + \beta_2 + \beta_4 x_{i4} + \epsilon_i$
- (Grupo C)  $y_i = \beta_0 + \beta_3 + \beta_4 x_{i4} + \epsilon_i$

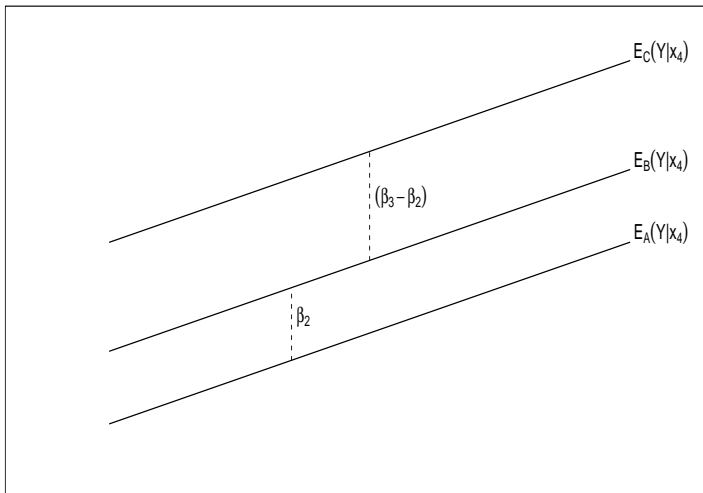
com diferenças de valores esperados

- $E_B(Y_i|x_{i4}) - E_A(Y_i|x_{i4}) = \beta_2$
- $E_C(Y_i|x_{i4}) - E_A(Y_i|x_{i4}) = \beta_3,$

para  $i = 1, \dots, n$ .

Assim, os efeitos  $\beta_2$  e  $\beta_3$  são incrementos nos valores esperados dos grupos B e C, respectivamente, com relação ao grupo A.

# Ilustração Ausência de Interação



## Ausência de Interação

Em forma matricial o modelo com ausência de interação fica dado por  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , em que  $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \mathbf{y}_3^\top)^\top$  com  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^\top$ , para  $i = 1, 2, 3$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_3, \beta_4)^\top$  e matriz  $\mathbf{X}$  dada por

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & x_{114} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & x_{1n_14} \\ 1 & 1 & 0 & x_{214} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & x_{2n_24} \\ 1 & 0 & 1 & x_{314} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & x_{3n_34} \end{bmatrix} .$$

## Presença de Interação

O modelo com interação entre a variável categórica  $X$  e a variável contínua  $X_4$  pode ser expresso na seguinte forma:

$$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i2} x_{i4} + \beta_6 x_{i3} x_{i4} + \epsilon_i,$$

em que  $y_1, \dots, y_n$  denotam os valores observados da variável resposta,  $x_{i2}$  e  $x_{i3}$  são valores de variáveis binárias representando os grupos B e C, respectivamente, enquanto  $x_{i4}$  representa os valores observados de uma variável contínua e  $\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$ , para  $i = 1, \dots, n$ .

## Presença de Interação

Portanto, tem-se três submodelos

- (Grupo A)  $y_i = \beta_0 + \beta_4 x_{i4} + \epsilon_i$
- (Grupo B)  $y_i = \beta_0 + \beta_2 + \beta_4 x_{i4} + \beta_5 x_{i4} + \epsilon_i$
- (Grupo C)  $y_i = \beta_0 + \beta_3 + \beta_4 x_{i4} + \beta_6 x_{i4} + \epsilon_i$

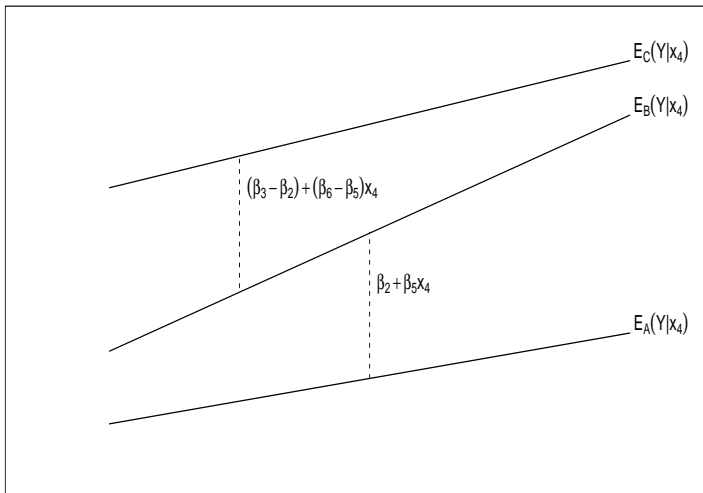
com diferenças de valores esperados

- $E_B(Y_i|x_{i4}) - E_A(Y_i|x_{i4}) = \beta_2 + \beta_5 x_{i4}$
- $E_C(Y_i|x_{i4}) - E_A(Y_i|x_{i4}) = \beta_3 + \beta_6 x_{i4},$

para  $i = 1, \dots, n.$

Assim, nota-se que as diferenças entre os valores esperados dependem dos valores da variável explicativa  $X_4$ .

# Ilustração Presença de Interação



## Presença de Interação

Em forma matricial o modelo com presença de interação fica dado por  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , em que  $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \mathbf{y}_3^\top)^\top$  com  $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})^\top$ , para  $i = 1, 2, 3$ ,  $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^\top$  e matriz  $\mathbf{X}$  dada por

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & x_{114} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & x_{1n_14} & 0 & 0 \\ 1 & 1 & 0 & x_{214} & x_{214} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & x_{2n_24} & x_{2n_24} & 0 \\ 1 & 0 & 1 & x_{314} & 0 & x_{314} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & x_{3n_34} & 0 & x_{3n_34} \end{bmatrix}.$$

- 1 Introdução
- 2 Variável Explicativa Binária
- 3 Variável Explicativa Categórica
- 4 Referências**



### Referência

- Montgomery, D. C.; Peck, E. A. e Vining, G. G. (2021). *Introduction to Linear Regression Analysis, 6th Edition*. Hoboken: Wiley.