

Modelos de Regressão

Paulo Roberto Borges de Souza Júnior

Laboratório de Informação em Saúde (LIS)
Programa de Pós-Graduação em Informação e Comunicação em Saúde (PPGICS)
Instituto de Comunicação e Informação Científica e Tecnológica (ICICT)
Fundação Oswaldo Cruz (Fiocruz)

Regressão Vs. Correlação

Muitas vezes, a relação entre duas variáveis pode ser explicada por uma função matemática. Ou seja, a magnitude de uma variável (dependente) é uma função da magnitude de outra variável (independente). Porém o contrário não é verdadeiro.

Regressão Vs. Correlação

Por exemplo:

- Pressão arterial (PA) X Idade -> em humanos
- O aumento da PA está relacionado ao aumento da idade, ou seja, a idade é um fator explicativo para o aumento da PA (não quer dizer que seja um fator determinante da PA);
- A PA é considerada como variável dependente e a idade é a variável independente, pois a PA não determina a idade;

Regressão Vs. Correlação

Esta relação de dependência é chamada de regressão

- Variável dependente -> também chamada de variável resposta
- Variável independente -> também chamada de variável preditora

Regressão Vs. Correlação

Em muitos casos, porém, a relação entre duas variáveis não é de dependência. Nestes casos, a magnitude de uma variável se modifica conforme modifica-se a magnitude da segunda variável, mas não é razoável considerar que existe uma variável dependente e outra independente.

Regressão Vs. Correlação

Por exemplo:

Comprimento dos braços X comprimento das pernas (em humanos)

Esta relação por ser descrita, porém, não há justificativa para afirmar-se que o comprimento de um membro depende do comprimento do outro.

Nestes casos, a análise da correlação é mais adequada que a análise de regressão.

Regressão

Tem por objetivo estabelecer uma função matemática que descreva a relação entre uma variável dependente (resposta) e uma ou mais variáveis independentes (preditoras)

$$y = f(x_1, x_2, \dots, x_k) + \varepsilon$$

Onde,

y é a variável dependente

x são as variáveis independentes

$f(x_1, x_2, \dots, x_k)$ é a função que descreve a variação sistemática

ε é a variação não sistemática (aleatória)

Regressão

A função f deve ser inferida a partir das observações das variáveis y, x_1, x_2, \dots, x_k .

Regressão Linear -> quando f pode ser representada por uma equação linear

Regressão Linear Simples

Tem por objetivo analisar a relação entre uma variável dependente (y) e uma única variável independente (x) através de uma equação linear:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

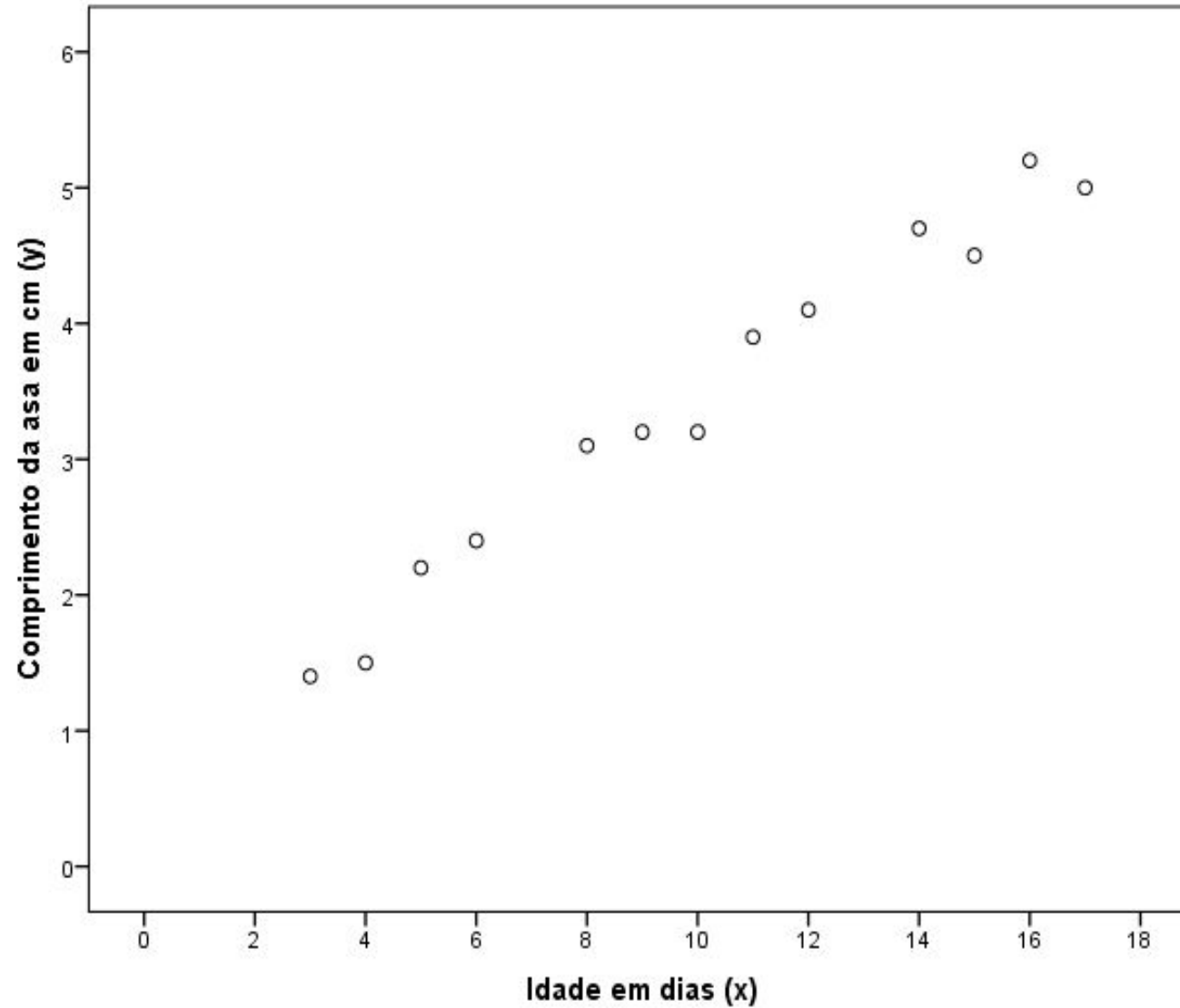
- β_0 e β_1 são constantes não conhecidas que serão estimadas a partir dos dados disponíveis
- Permite prever a variável y a partir das observações de x

Regressão Linear Simples

- A equação da regressão linear simples

Usando como exemplo os dados sobre comprimento das asas e idade de 13 pardais (Zar, 1999), temos o seguinte diagrama de dispersão:

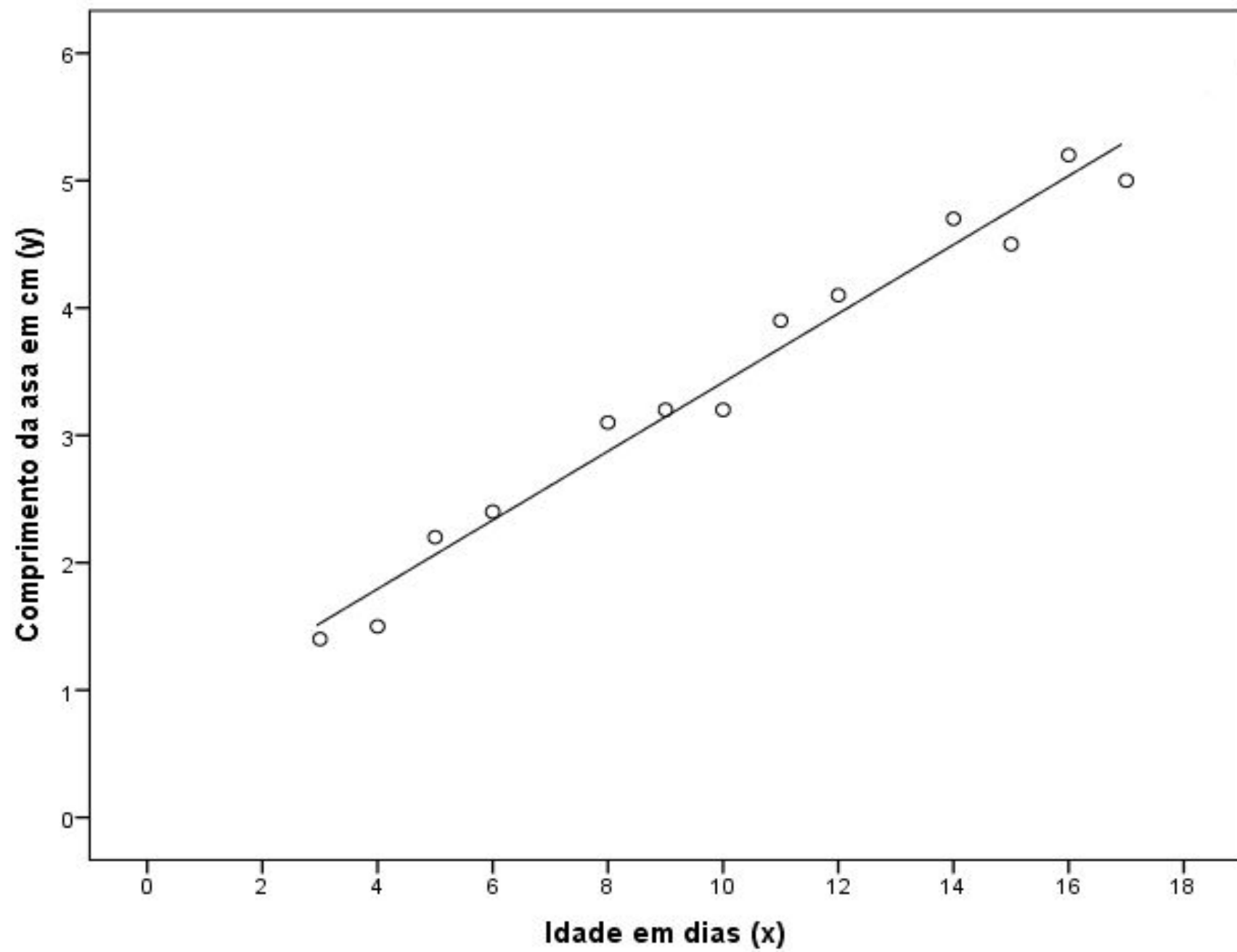
Regressão Linear Simples

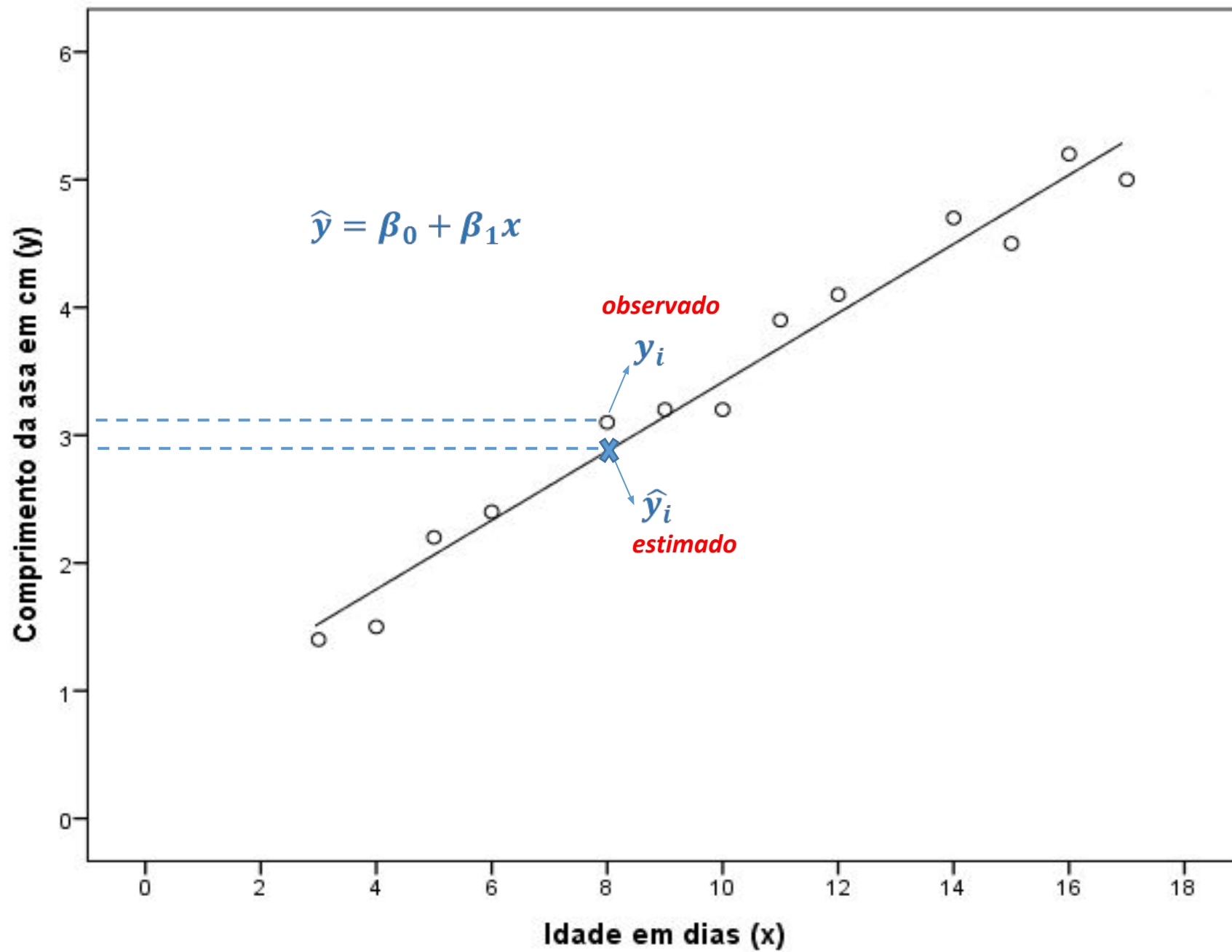


Regressão Linear Simples

Cada ponto do gráfico representa um par de valores de x e y , ou seja $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_{13}, y_{13})$.

A simples observação do gráfico permite imaginar que poderíamos traçar uma reta entre os pontos.





Regressão Linear Simples

Qual a melhor reta?

Em uma situação em que todos os pontos de um diagrama de dispersão se encontrassem em uma linha reta, não teríamos que nos preocupar em encontrar a reta que melhor resume os pontos do diagrama.

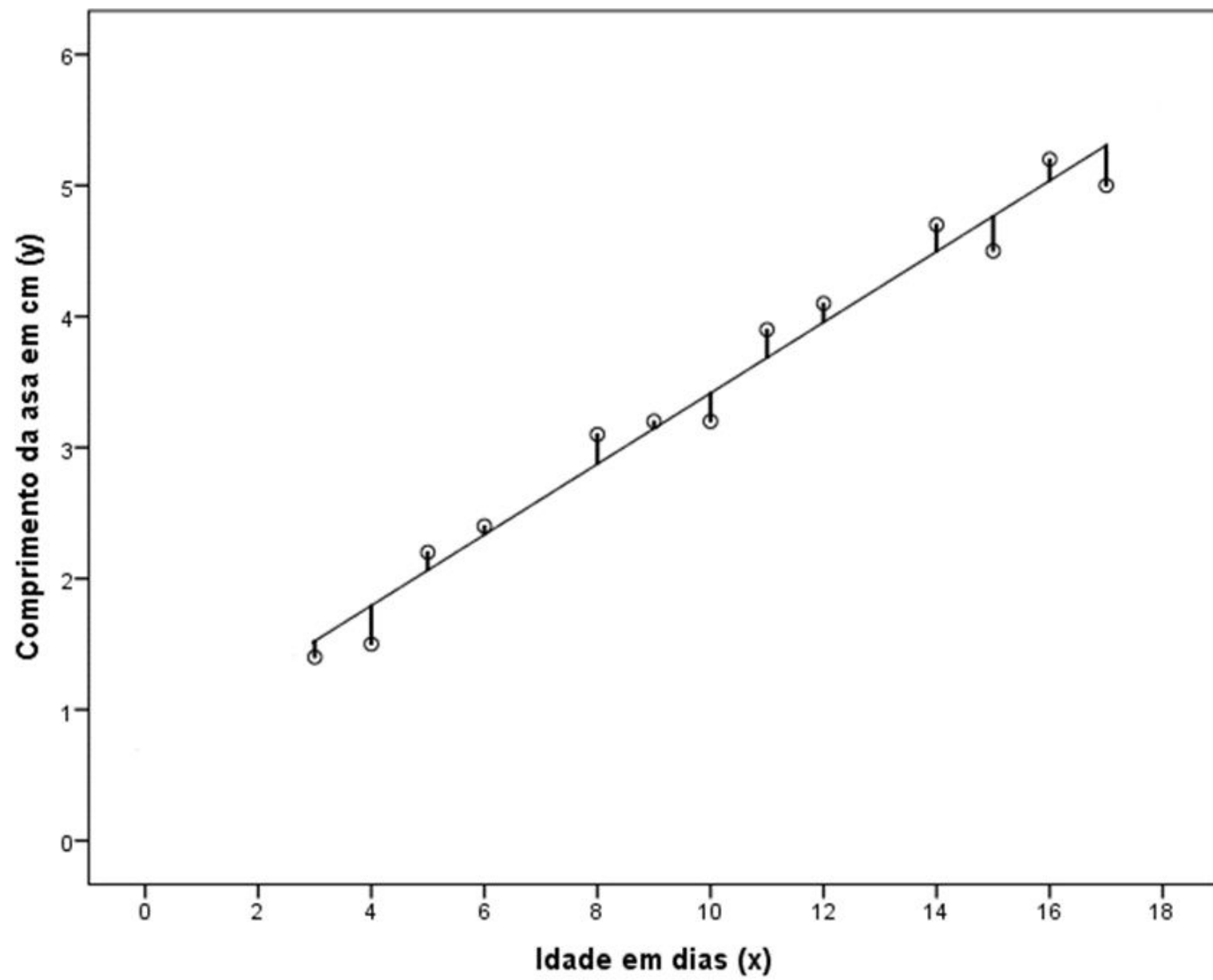
Porém, em uma nuvem de pontos mais realista, é possível traçar várias retas diferentes, cujos coeficientes β_0 e β_1 da equação são estimados por métodos distintos.

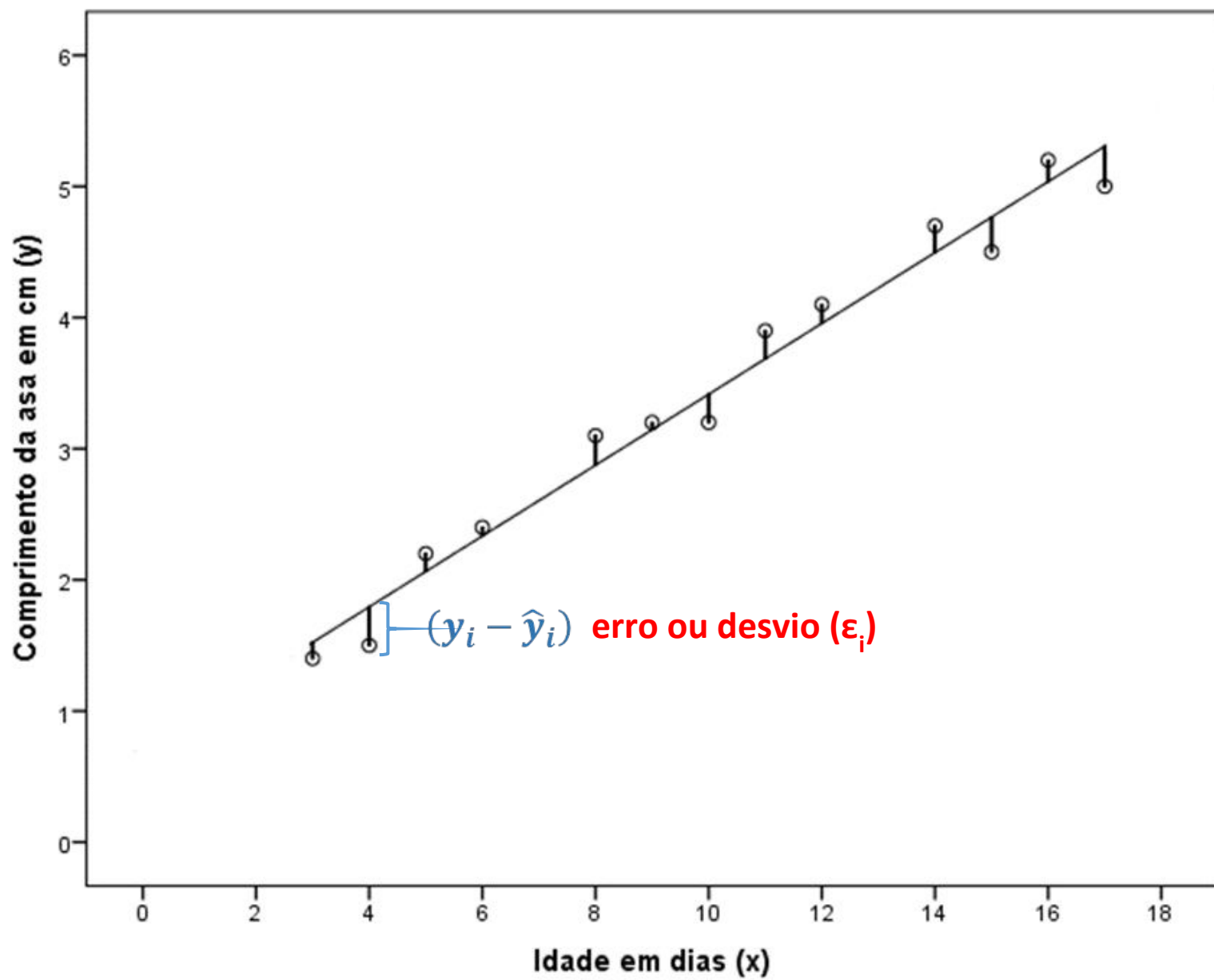
Regressão Linear Simples

- Entretanto, o método mais utilizado é o método dos mínimos quadrados (MQ), que consiste em encontrar a reta que minimiza a soma dos quadrados das distâncias verticais entre cada ponto e a reta.

Existe uma única reta cuja distâncias verticais quadráticas sejam mínimas.

A melhor reta pelo método MQ é aquela em que $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ é mínimo.





Regressão Linear Simples

O erro ε pode representar:

- Erro de medição:
 - Ex.: A medida do comprimento das asas dos pardais pode ter sido feita por profissionais diferentes, com instrumentos de medidas diferentes, com métodos diferentes, etc...
- Erro aleatório (estocástico): Ocorre devido a inerente “irreprodutibilidade” de fenômenos biológicos e sociais.
 - Mesmo se não houvesse erros de medição, a reprodução contínua de um experimento usando pardais da mesma faixa de idade, resultaria em comprimento de asas diferentes. Essas diferenças não são previsíveis e são chamadas de aleatórias.
 - Outros fatores que afetam a variável dependente Y , mas que não estão contempladas nas variáveis explicativas X .
- Forma funcional inadequada, por exemplo,

$$y = \beta_0 + \beta_1 x \quad \text{ou} \quad y = \beta_0 + \beta_1 x + \beta_1 x^2?$$

Hipóteses assumidas pelo modelo

H1 - Existência: Para qualquer valor de X , existe na população uma distribuição Normal dos valores de Y . Também significa que para cada valor de X , existe na população uma distribuição Normal dos ε 's.

H2 - Linearidade: A relação entre as variáveis é linear $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n$;

A variável dependente é a soma de um conjunto de elementos: a origem da reta (intercepto), uma combinação linear de variáveis independentes ou preditoras e os resíduos.

O não cumprimento deste pressuposto é um erro de especificação e pode ocorrer, por exemplo, devido a omissão de variáveis independentes importantes ou inclusão de variáveis independentes irrelevantes; a relação entre as variáveis dependente e independente não é linear.

Hipóteses assumidas pelo modelo

H3 - Independência: Os valores da variável aleatória Y (assim como dos resíduos) são independentes;

Os valores de y_2 não é influenciado pelos valor de y_1 , bem como o valor de ε_2 não é influenciado pelo valor de ε_1 , ou seja, $\text{Cov}(\varepsilon_i, \varepsilon_k) = 0$ para todo $i \neq k$.

H4 - Homocedasticidade: Para cada valor da variável independente (ou conjunto de variáveis independentes), a variância dos resíduos é constante;

$$V(\varepsilon_i) = \sigma^2 \quad \text{para todo } i=1,n.$$

Hipóteses assumidas pelo modelo

H5 - Normalidade: Para cada valor da variável independente (ou conjunto de variáveis independentes), os resíduos se distribuem normalmente e com média zero.;

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{para todo } i=1, n$$

ε_i são independentes e identicamente distribuídos $N(0, \sigma^2)$.

H6 – Ausência de colinearidade ou multicolinearidade: Não existe relação linear exata entre nenhuma das variáveis independentes. Este pressuposto é válido somente para modelos de regressão múltipla (com duas ou mais variáveis independentes);

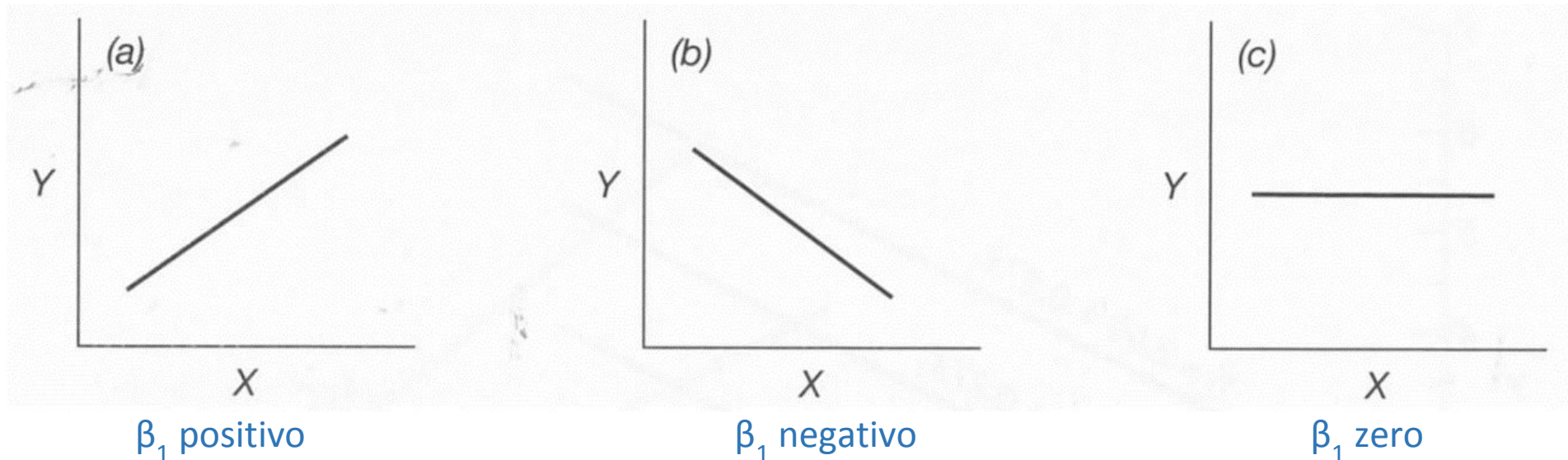
As colunas da matriz X são linearmente independentes.

Regressão Linear Simples

Os coeficientes β_0 e β_1

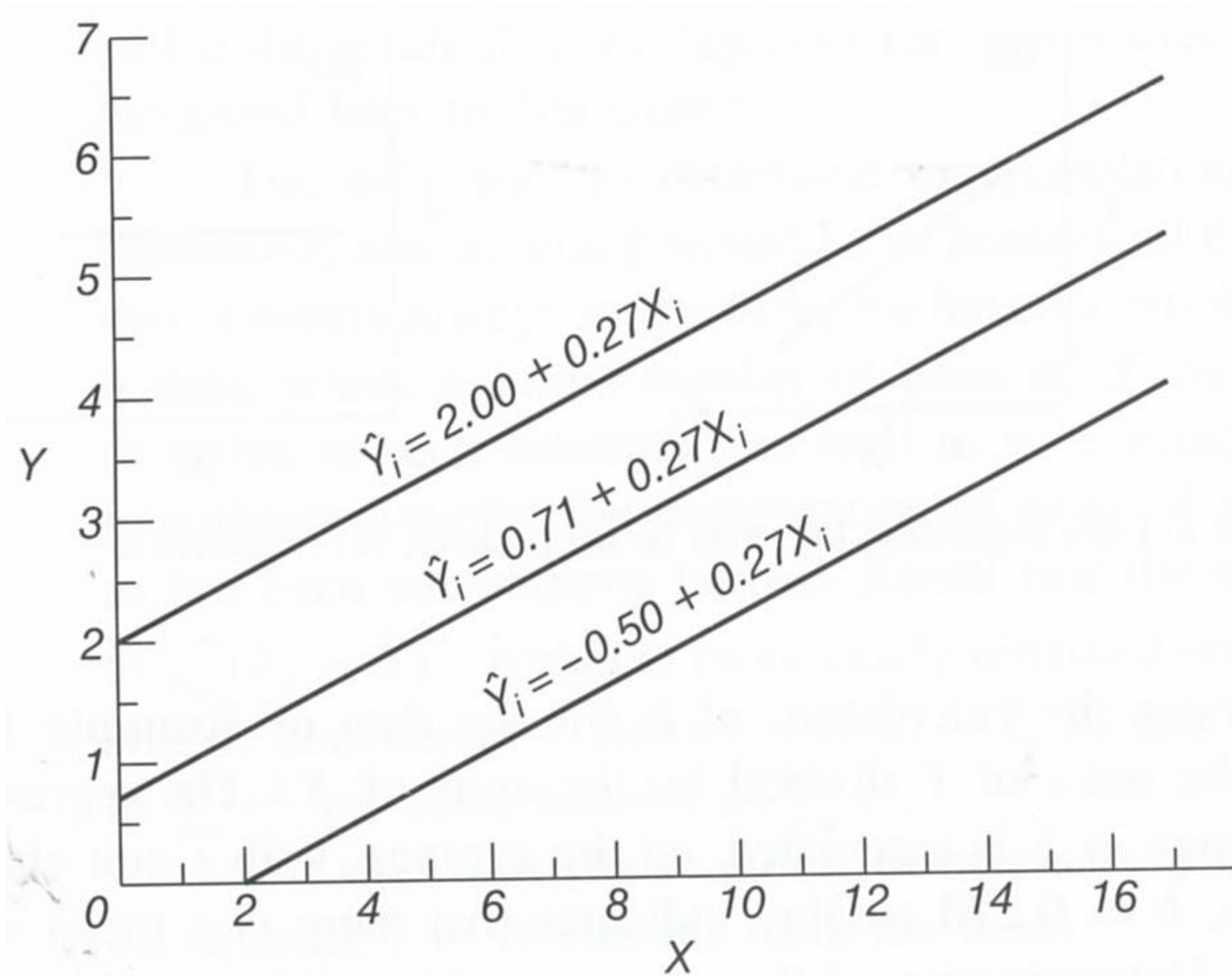
O parâmetro β_1 é o coeficiente angular da reta de regressão e define a inclinação da reta.

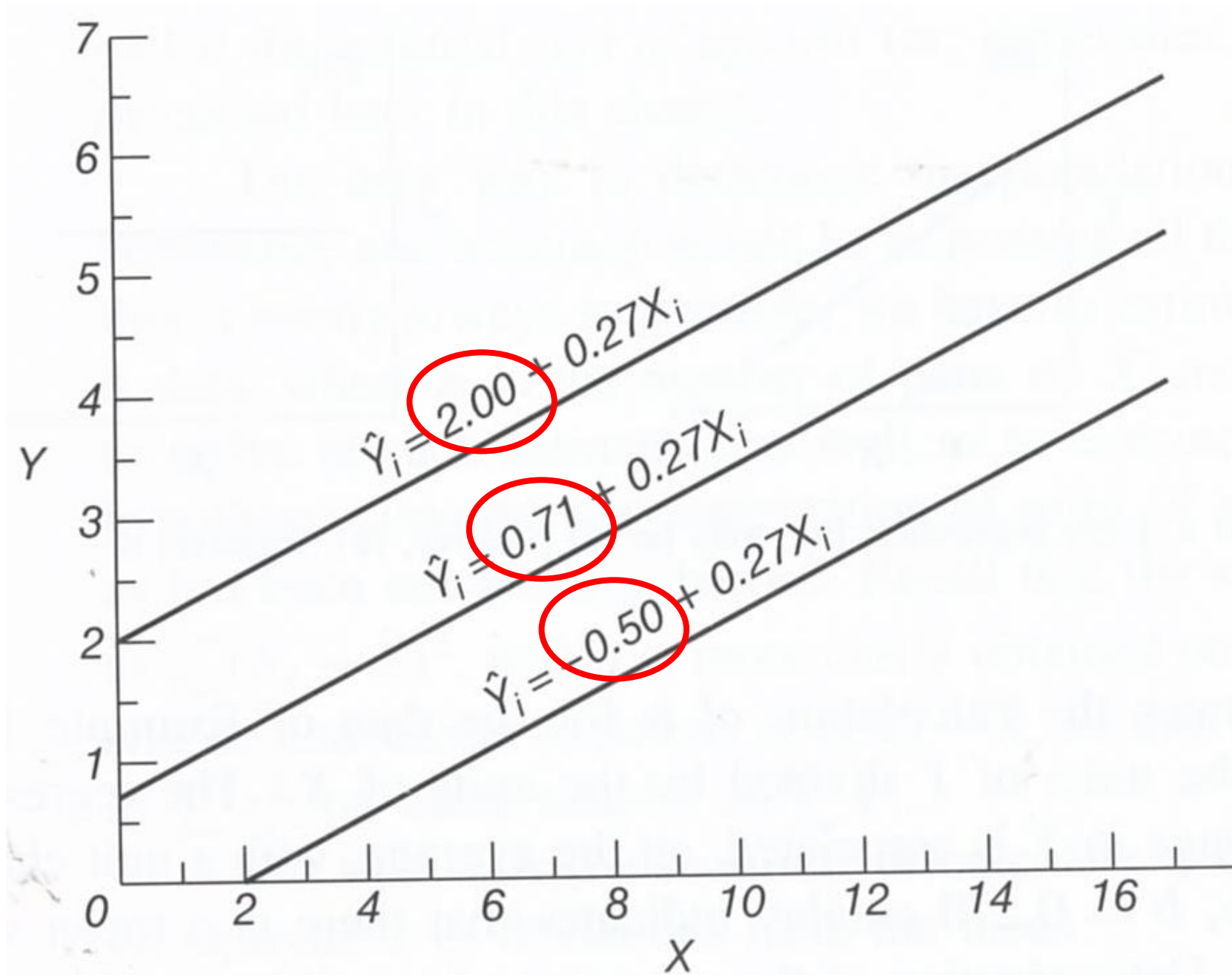
O valor de β_1 pode ser positivo ou negativo, variando, teoricamente entre $-\infty$ e $+\infty$, incluindo o zero.

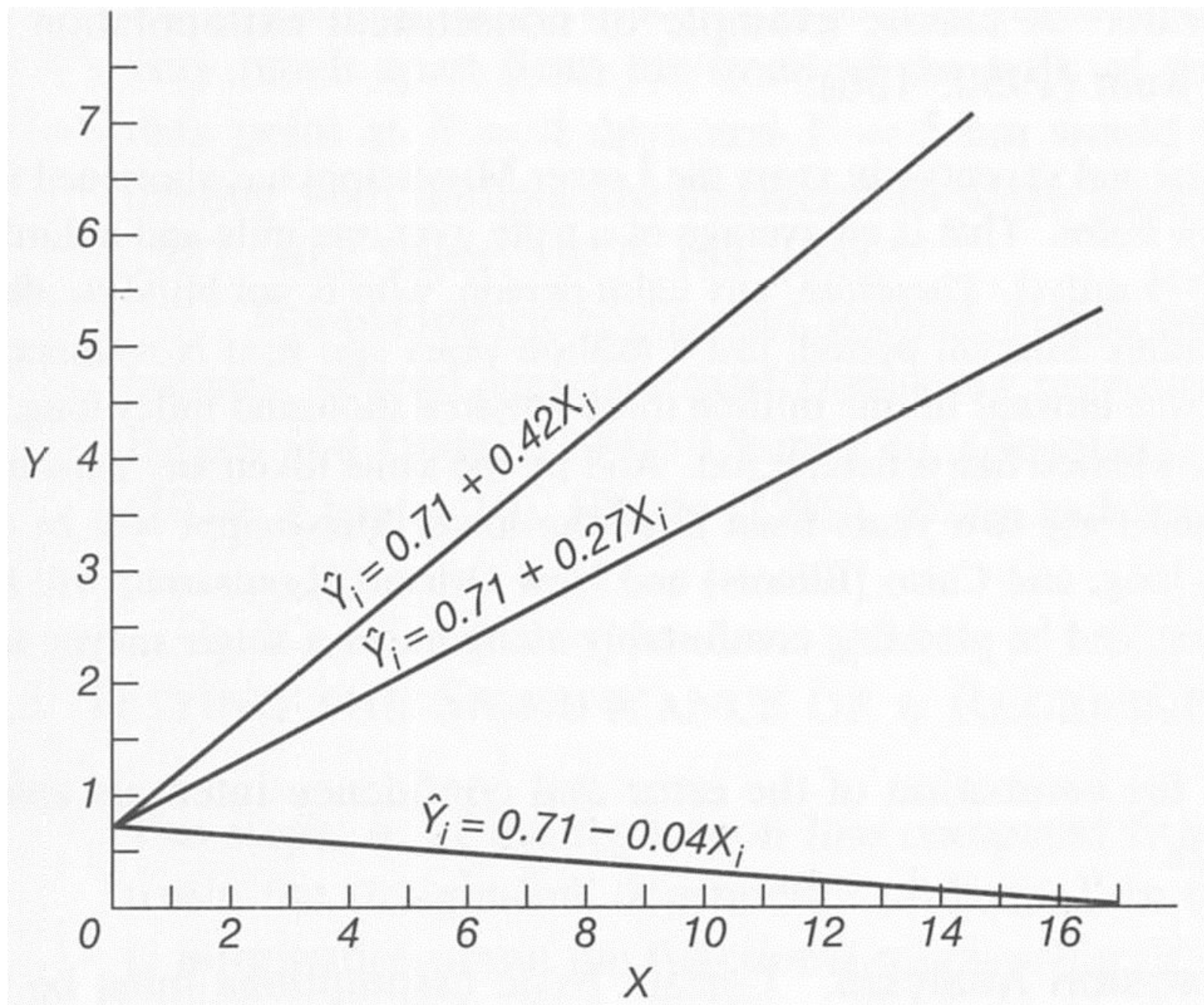


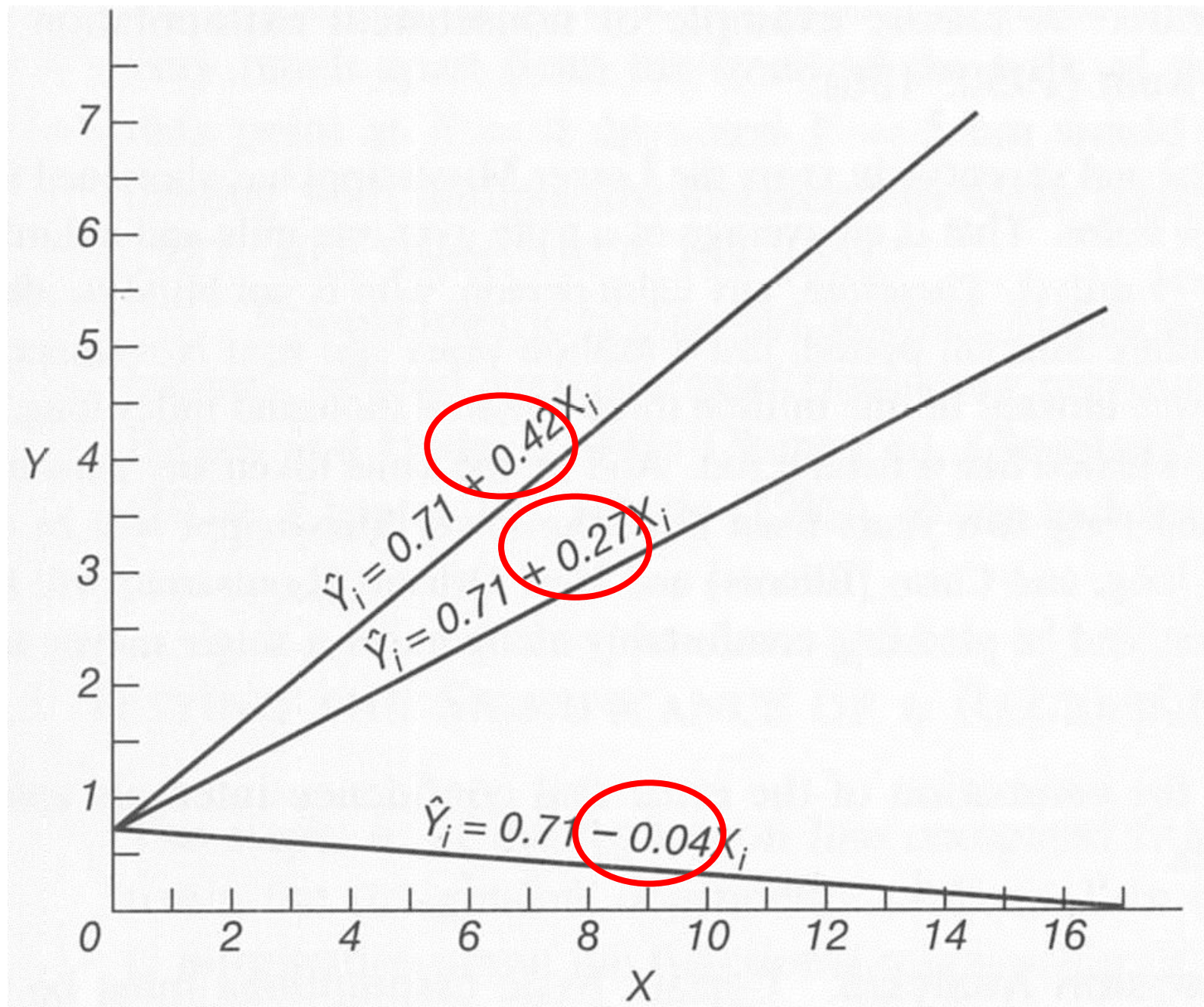
Regressão Linear Simples

Os coeficientes β_0 e β_1









Regressão Linear Simples

- Calculando os valores preditos de Y

cm

cm/dia

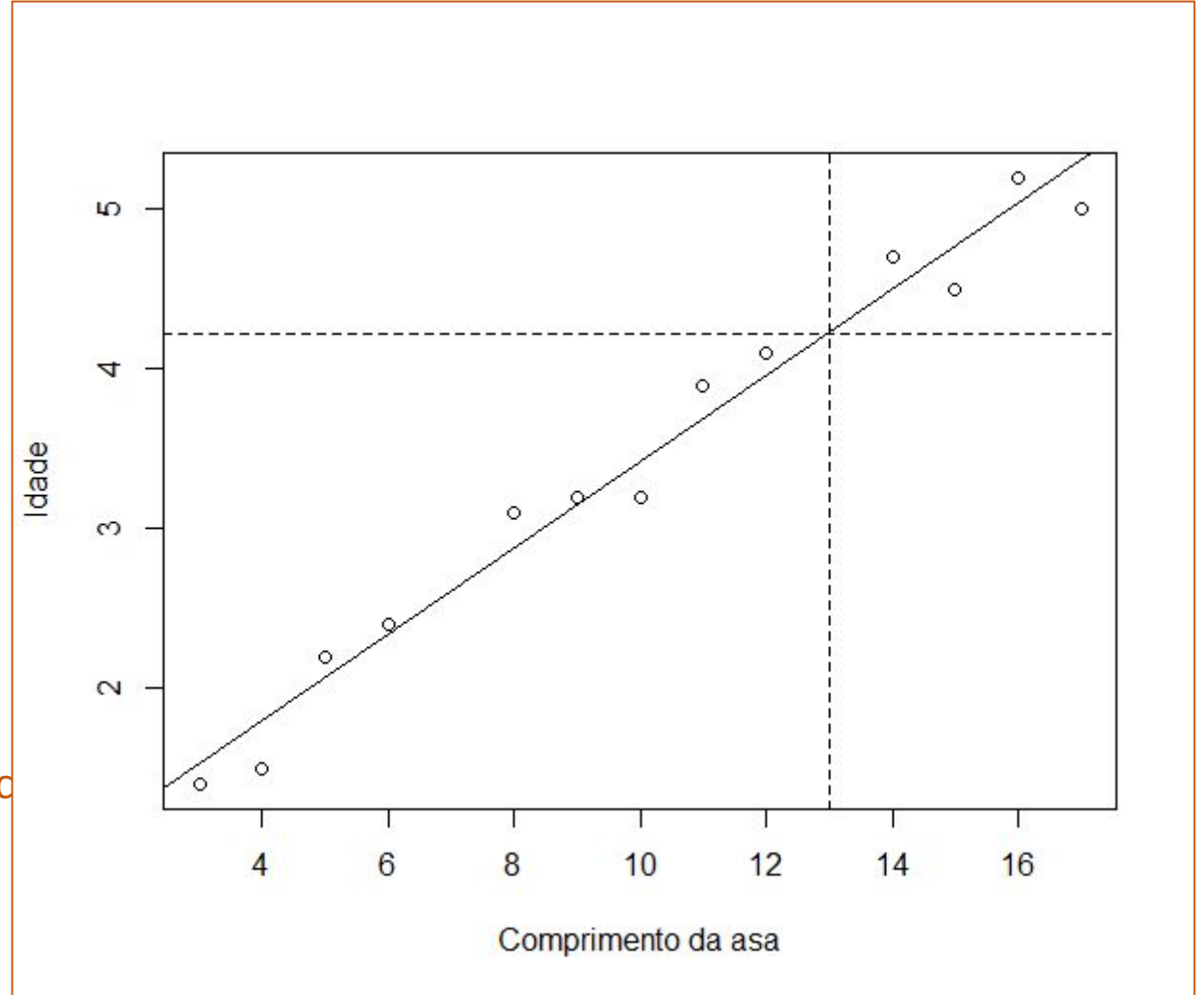
dias

Regressão Linear Simples

- Calculando os valores preditos de Y

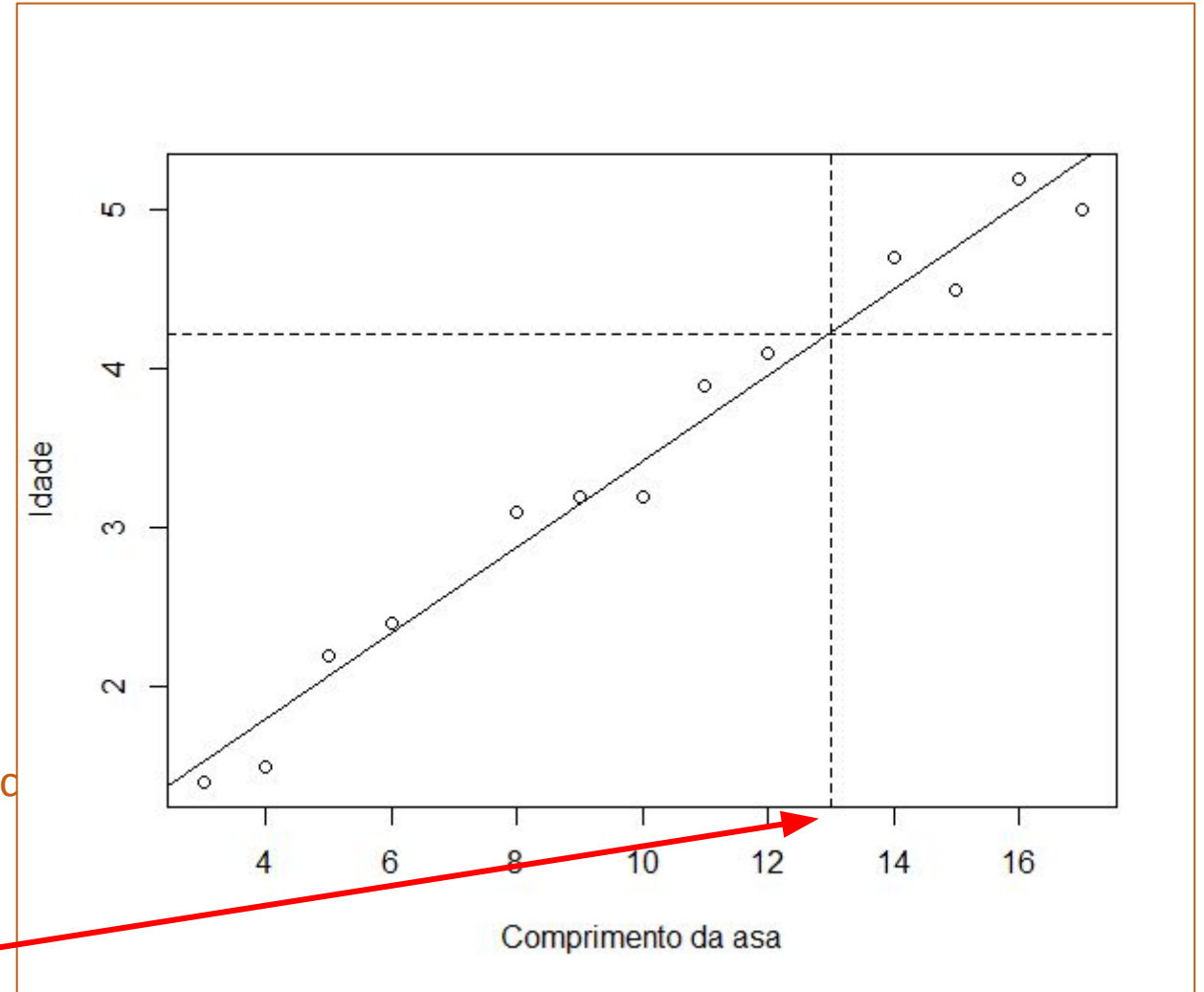
cm

o

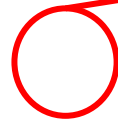


Regressão Linear Simples

- Calculando os valores preditos de Y

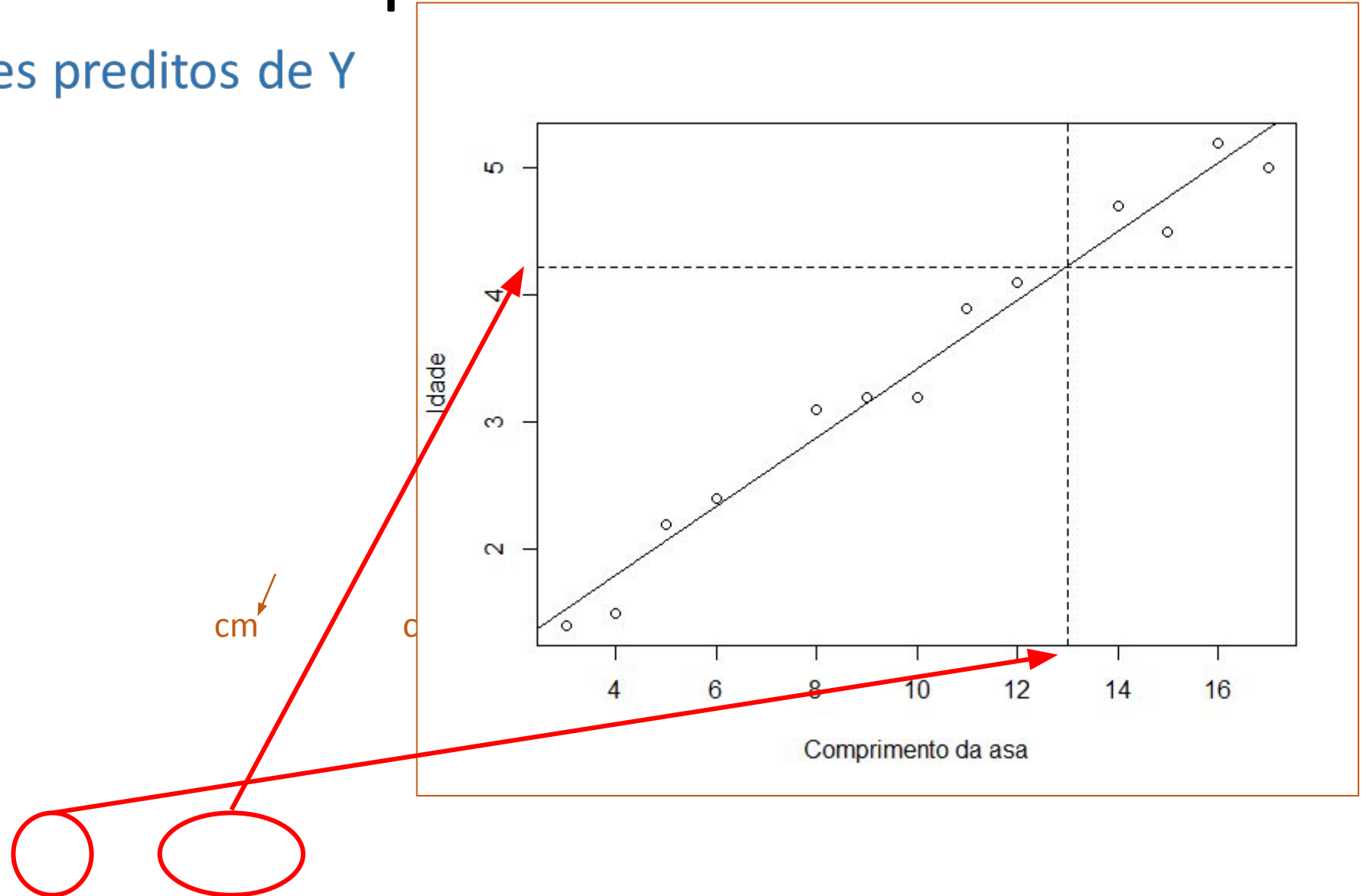


cm



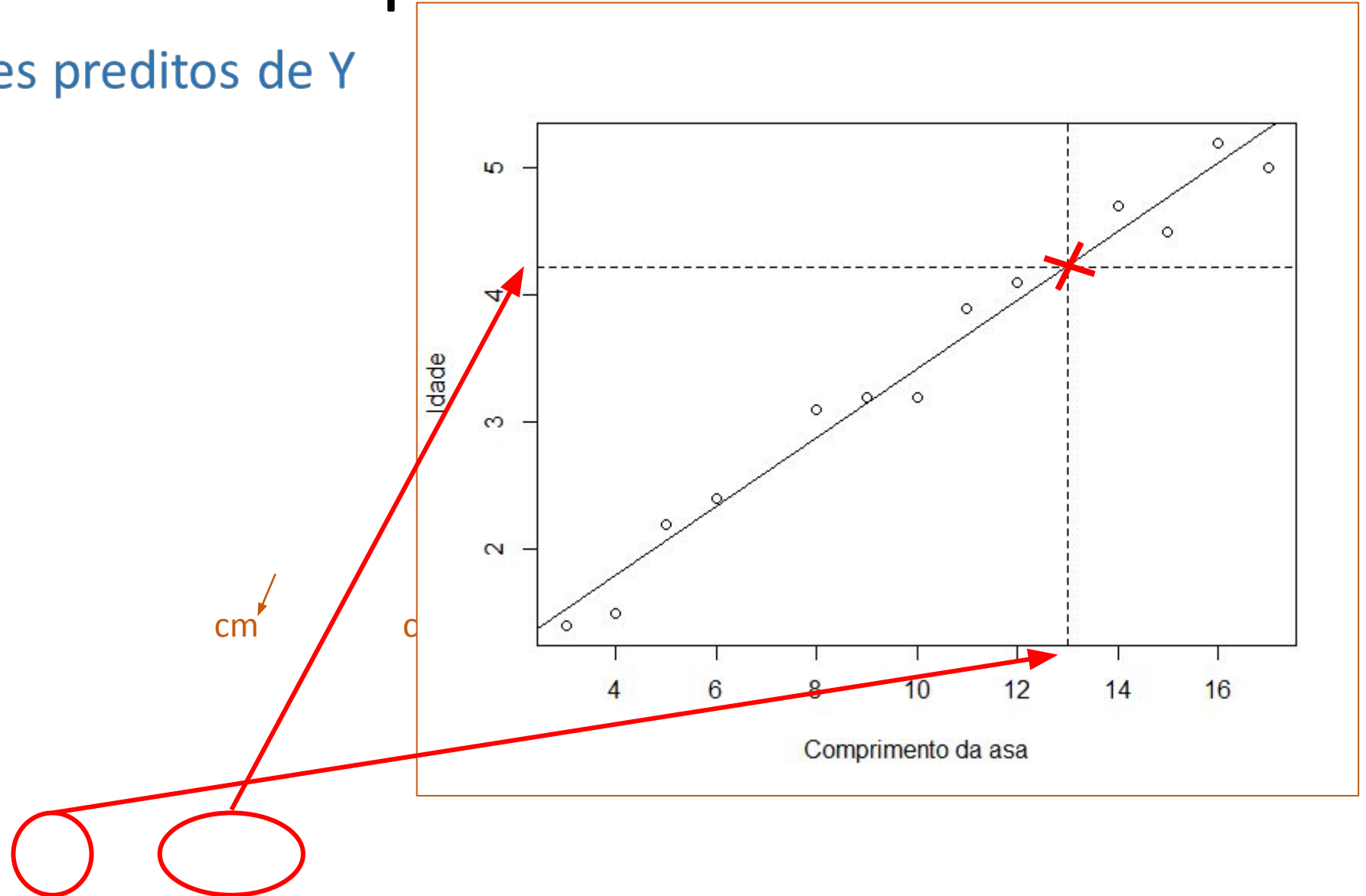
Regressão Linear Simples

- Calculando os valores preditos de Y



Regressão Linear Simples

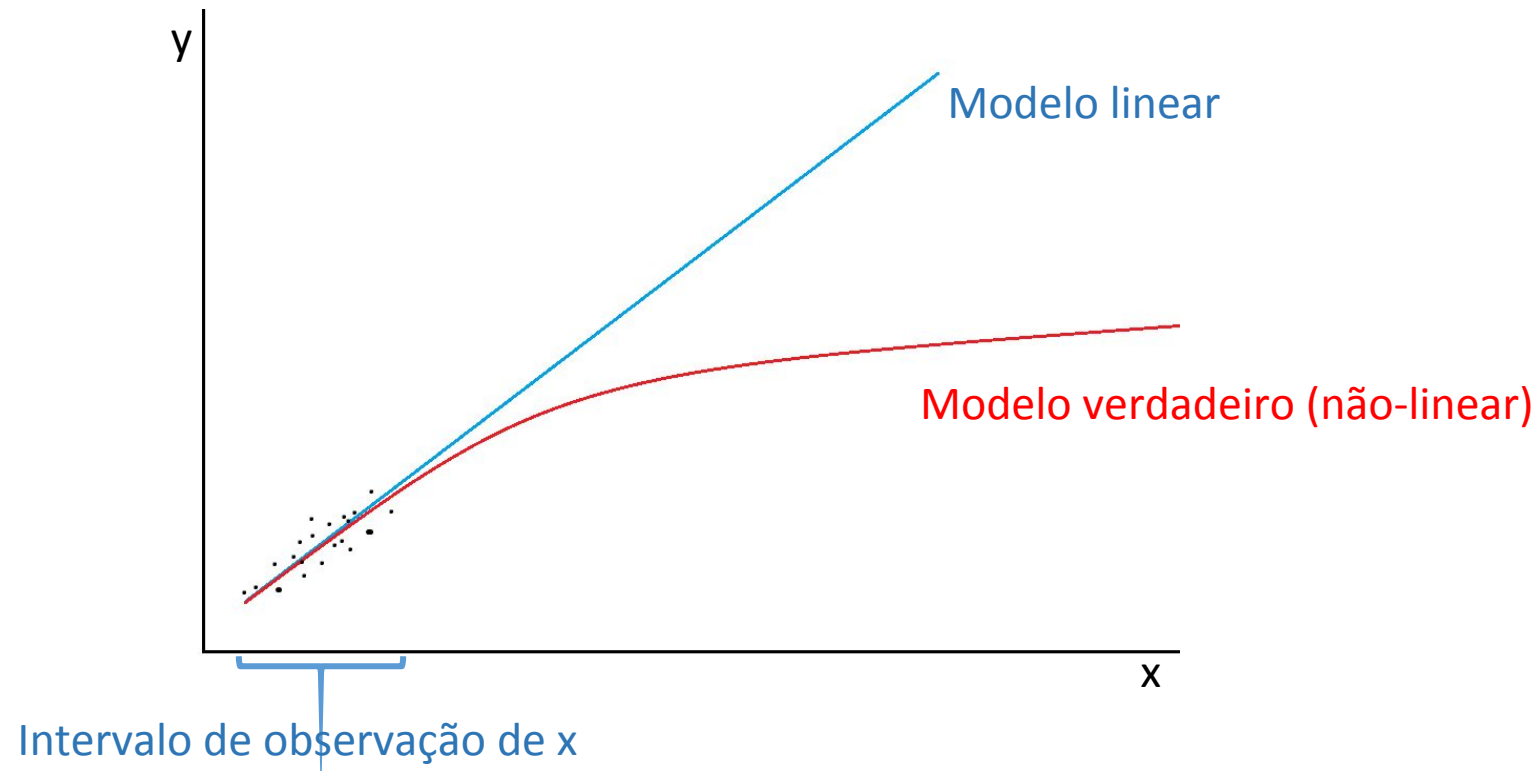
- Calculando os valores preditos de Y



Regressão Linear Simples

Cuidados que devemos ter para estimar os valores de Y

Os valores preditos de Y devem ser calculados dentro da faixa de valores observados de x, pois só conhecemos a função que relaciona y e x neste intervalo.



Regressão Linear Simples

Coeficiente de determinação

$$R^2 = \frac{SQ_{regressão}}{SQ_{total}}$$

Teste t de Student

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

$$t = \frac{\beta_1}{S_{\beta_1}}$$

$$t \sim t_{\alpha, n-2 \text{ gl}}$$

Regressão Linear Simples

Análise de variância (Exemplo dos pardais)

Fonte	Soma dos quadrados (SQ)	Graus de Liberdade (g.l.)	Quadrados Médios (QM)	F	P_valor
Regressão	19,132	1	19,132	401,087	<0,001
Erro	0,525	11	0,048		
Total	19,657	12			

Regressão Linear Simples

Análise de variância (Exemplo dos pardais)

Fonte	Soma dos quadrados (SQ)	Graus de Liberdade (g.l.)	Quadrados Médios (QM)	F	P_valor
Regressão	19,132	1	19,132	401,087	<0,001
Erro	0,525	11	0,048		
Total	19,657	12			

$$R^2 = \frac{19,132}{19,657} = 0,973 = 97,3\%$$

Regressão Linear Simples

Análise de variância (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
```

```
> anova(mod_exempl)
```

Analysis of Variance Table

Response: Comprimento da asa

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Idade	1	19.1322	19.1322	401.09	5.267e-10 ***
Residuals	11	0.5247	0.0477		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Regressão Linear Simples

Coeficientes do modelo (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
```

```
> summary(mod_exempl)
```

Call:

```
lm(formula = `Comprimento da asa` ~ Idade, data = pardais)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.30699	-0.21538	0.06553	0.16324	0.22507

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.71309	0.14790	4.821	0.000535 ***
Idade	0.27023	0.01349	20.027	5.27e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2184 on 11 degrees of freedom

Multiple R-squared: 0.9733, Adjusted R-squared: 0.9709

F-statistic: 401.1 on 1 and 11 DF, p-value: 5.267e-10

Regressão Linear Simples

Coeficientes do modelo (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
```

```
> summary(mod_exempl)
```

Call:

```
lm(formula = `Comprimento da asa` ~ Idade, data = pardais)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.30699	-0.21538	0.06553	0.16324	0.22507

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.71309	0.14790	4.821	0.000535 ***
Idade	0.27023	0.01349	20.027	5.27e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2184 on 11 degrees of freedom
Multiple R-squared: 0.9733, Adjusted R-squared: 0.9709
F-statistic: 401.1 on 1 and 11 DF, p-value: 5.267e-10

β_0

β_1

Regressão Linear Simples

Coeficientes do modelo (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
```

```
> summary(mod_exempl)
```

Call:

lm(formula = `Comprimento da asa` ~ Idade, data = pardais)

Residuals:

Min	1Q	Median	3Q	Max
-0.30699	-0.21538	0.06553	0.16324	0.22507

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.71309	0.14790	4.821	0.000535 ***
Idade	0.27023	0.01349	20.027	5.27e-10 ***

Signif. levels: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2184 on 11 degrees of freedom
Multiple R-squared: 0.9733, Adjusted R-squared: 0.9709
F-statistic: 401.1 on 1 and 11 DF, p-value: 5.267e-10

β_0

S_{β_0}

β_1

S_{β_1}

Regressão Linear Simples

Coeficientes do modelo (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
```

```
> summary(mod_exempl)
```

Call:

lm(formula = `Comprimento da asa` ~ Idade, data = pardais)

Residuals:

Min	1Q	Median	3Q	Max
-0.30699	-0.21538	0.06553	0.16324	0.22507

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.71309	0.14790	4.821	0.000535 ***
Idade	0.27023	0.01349	20.027	5.27e-10 ***

Signif. levels: '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2184 on 11 degrees of freedom
Multiple R-squared: 0.9733, Adjusted R-squared: 0.9709
F-statistic: 401.1 on 1 and 11 DF, p-value: 5.267e-10

$$t = \frac{\beta_0}{s_{\beta_0}} = \frac{0,713}{0,148} = 4,82$$

$$t = \frac{\beta_1}{s_{\beta_1}} = \frac{0,270}{0,013} = 20,027$$

Regressão Linear Simples

Coeficientes do modelo (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
```

```
> summary(mod_exempl)
```

Call:

lm(formula = `Comprimento da asa` ~ Idade, data = pardais)

Residuals:

Min	1Q	Median	3Q	Max
-0.30699	-0.21538	0.06553	0.16324	0.22507

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.71309	0.14790	4.821	0.000535 ***
Idade	0.27023	0.01349	20.027	5.27e-10 ***

Signif.

*** 0.001 ** 0.01 * 0.05 . 0.1 ' 1

Residual standard error: 0.2184 on 11 degrees of freedom
Multiple R-squared: 0.9733, Adjusted R-squared: 0.9709
F-statistic: 401.1 on 1 and 11 DF, p-value: 5.267e-10

$$t = \frac{\beta_0}{s_{\beta_0}} = \frac{0,713}{0,148} = 4,82$$

$$t \sim t_{0,05; 11gl}$$

$$t = \frac{\beta_1}{s_{\beta_1}} = \frac{0,270}{0,013} = 20,027$$

Regressão Linear Simples

Coeficientes do modelo (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
```

```
> summary(mod_exempl)
```

Call:

```
lm(formula = `Comprimento da asa` ~ Idade, data = pardais)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.30699	-0.21538	0.06553	0.16324	0.22507

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.71309	0.14790	4.821	0.000535 ***
Idade	0.27023	0.01349	20.027	5.27e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2184 on 11 degrees of freedom

Multiple R-squared: 0.9733. Adjusted R-squared: 0.9709

F-statistic: 401.1 on 1 and 11 DF, p-value: 5.267e-10

$$R^2 = \frac{19,132}{19,657} = 0,973 = 97,3\%$$

Regressão Linear Simples

Coeficientes do modelo (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
```

```
> summary(mod_exempl)
```

Call:

```
lm(formula = `Comprimento da asa` ~ Idade, data = pardais)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.30699	-0.21538	0.06553	0.16324	0.22507

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.71309	0.14790	4.821	0.000535 ***
Idade	0.27023	0.01349	20.027	5.27e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2184 on 11 degrees of freedom

Multiple R-squared: 0.9733, Adjusted R-squared: 0.9709

F-statistic: 401.1 on 1 and 11 DF, p-value: 5.267e-10

Ajustado pelos graus de liberdade do modelo, ou seja, pelo número de variáveis independente (k)

$$R_a^2 = 1 - \frac{\frac{SQ_{residual}}{n - k - 1}}{\frac{SQ_{total}}{n - 1}}$$

$$R_a^2 = 1 - \frac{\frac{0,525}{13 - 1 - 1}}{\frac{19,657}{13 - 1}}$$

Regressão Linear Múltipla

Regressão Linear Múltipla

- É uma extensão da regressão linear simples
- Descreve a relação entre uma variável dependente (Y) e duas ou mais variáveis independentes (X_1, X_2, \dots, X_k)
- Mede o efeito conjunto das variáveis independentes na variável dependente
- Permite identificar quais variáveis independentes são mais importantes na predição da variável dependente

Regressão Linear Múltipla

- Mesmo se o interesse for em estimar o efeito de uma única variável independente sobre Y , geralmente, é interessante incluir outras variáveis que influenciam Y em uma análise de regressão múltipla, por dois motivos:
 - Reduzir o erro estocástico e, portanto, reduzir a variância do resíduo (ϵ). Isso torna as estimativas mais precisas
 - Eliminar o viés que pode resultar ao ignorarmos uma variável que afeta substancialmente Y

Regressão Linear Múltipla

Dificuldades

- É difícil escolher o melhor modelo, pois podem existir várias variáveis independentes para serem utilizadas (testadas) no modelo
- A representação gráfica do modelo ajustado torna-se muito complicada quando o modelo apresenta mais de 3 variáveis (3 dimensões)
- O cálculo e a interpretação são mais complicados

Regressão Linear Múltipla

Modelo estatístico

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

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$ são parâmetros do modelo

- Uma variável dependente contínua (Y)
- Duas ou mais variáveis independentes quantitativas ou qualitativas (dummy)

Regressão Linear Múltipla

Modelo estatístico

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

$\beta_0, \beta_1, \beta_2, \dots, \beta_k$ são parâmetros do modelo

- A equação de regressão no modelo multivariado não define uma reta em um plano, como no modelo simples, e sim um hiperplano em um espaço multidimensional ($k+1$ dimensões)

Regressão Linear Múltipla

- Cada parâmetro β_i ($i=1,\dots,k$) da regressão pode ser interpretado da seguinte forma:
 - variação na resposta média, $E(Y)$, associada a variação unitária em X_i , controlando o efeito das outras variáveis
 - variação na resposta média, $E(Y)$, associada a variação unitária em X_i , mantendo constante as outras variáveis
 - efeito parcial de cada variável na variável resposta
- β_0 : valor esperado de Y quando todas as outras variáveis independentes são iguais a zero.

Regressão Linear Múltipla

Estimando os coeficientes

Os coeficientes da equação de regressão múltipla também são estimados pelo Método dos Mínimos Quadrados (MQ), ou seja, os coeficientes estimados são aqueles que minimizam a soma dos quadrados dos resíduos:

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})]^2$$

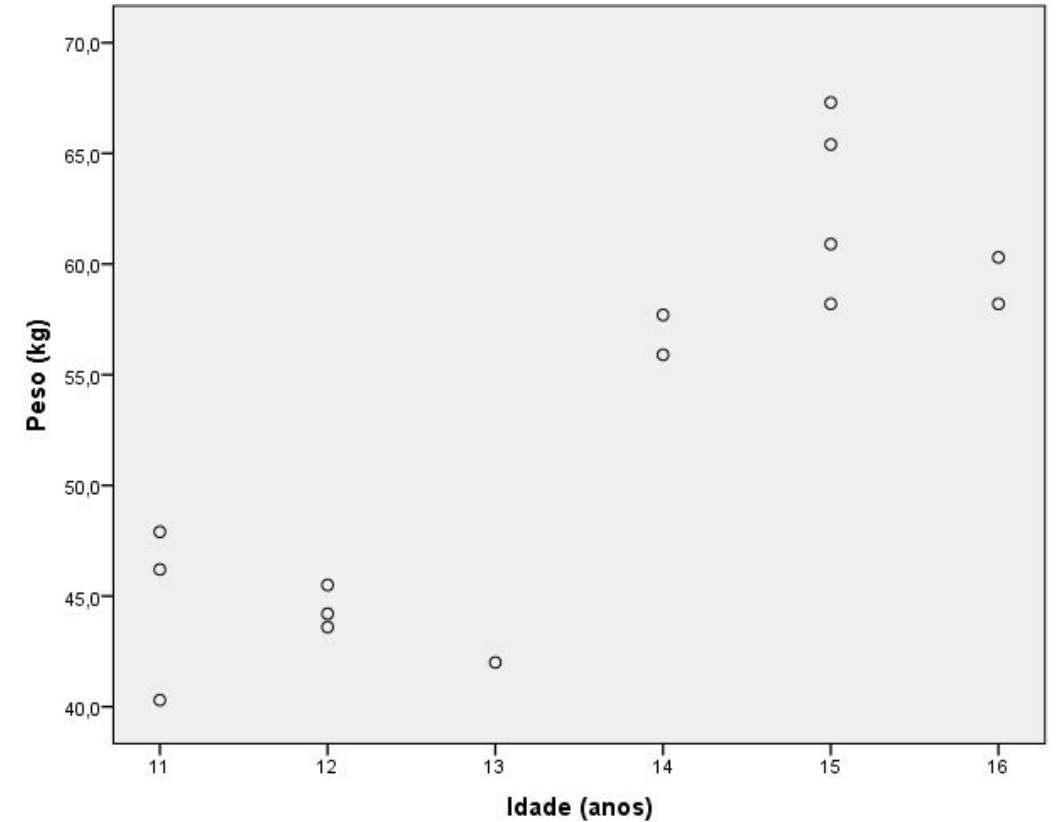
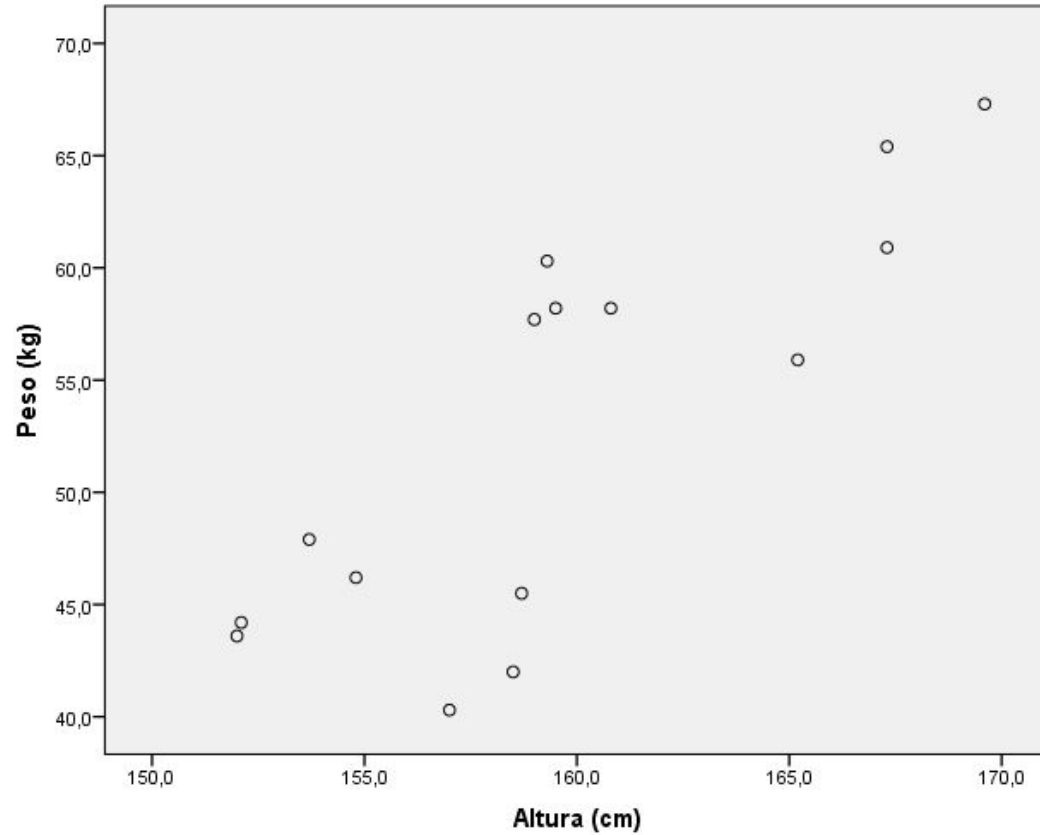
Regressão Linear Múltipla

Exemplo: Dados de peso, idade e altura de 15 crianças:

Criança	Peso (kg)	Altura (cm)	Idade (anos)
1	40.3	157.0	11
2	42.0	158.5	13
3	43.6	152.0	12
4	44.2	152.1	12
5	45.5	158.7	12
6	47.9	153.7	11
7	46.2	154.8	11
8	55.9	165.2	14
9	57.7	159.0	14
10	58.2	159.5	15
11	60.9	167.3	15
12	60.3	159.3	16
13	65.4	167.3	15
14	67.3	169.6	15
15	58.2	160.8	16

Regressão Linear Múltipla

Exemplo: Dados de peso, idade e altura de 15 crianças:



Regressão Linear Múltipla

A equação de regressão

$$E(Y_i | X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

define um plano passando pelo meio da nuvem de pontos.

Este plano representa o valor esperado do peso em função da altura e da idade das crianças.

Regressão Linear Múltipla

Estatísticas descritivas

Estatísticas descritivas

	Média	Desvio padrão	N
Peso (kg)	52,907	9,0039	15
Altura (cm)	159,653	5,5577	15
Idade (anos)	13,47	1,846	15

Regressão Linear Múltipla

Ajuste do modelo

```
> mod_exempl2 <- lm(Peso ~ Altura + Idade, data=criancas)
> summary(mod_exempl2)
```

Call: `lm(formula = Peso ~ Altura + Idade, data = criancas)`

Residuals:

Min	1Q	Median	3Q	Max
-8.8356	-2.3674	0.4124	3.2703	5.8455

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-90.8133	36.7238	-2.473	0.02934 *
Altura	0.6643	0.2730	2.433	0.03155 *
Idade	2.7961	0.8218	3.402	0.00525 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.056 on 12 degrees of freedom

Multiple R-squared: 0.826, Adjusted R-squared: 0.797

F-statistic: 28.49 on 2 and 12 DF, p-value: 2.773e-05

Regressão Linear Múltipla

Ajuste do modelo

```
> mod_exempl2 <- lm(Peso ~ Altura + Idade, data=criancas)
> summary(mod_exempl2)
```

Call: `lm(formula = Peso ~ Altura + Idade, data = criancas)`

Residuals:

Min	1Q	Median	3Q	Max
-8.8356	-2.3674	0.4124	3.2703	5.8455

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-90.8133	36.7238	-2.473	0.02934 *
Altura	0.6643	0.2730	2.433	0.03155 *
Idade	2.7961	0.8218	3.402	0.00525 **

$$Y = -90,813 + 0,664*(Altura) + 2,796*(Idade)$$

Residual standard error: 4.056 on 12 degrees of freedom

Multiple R-squared: 0.826, Adjusted R-squared: 0.797

F-statistic: 28.49 on 2 and 12 DF, p-value: 2.773e-05

Regressão Linear Múltipla

Ajuste do modelo

```
> mod_exempl2 <- lm(Peso ~ Altura + Idade, data=criancas)
```

```
> summary(mod_exempl2)
```

```
Call: lm(formula = Peso ~ Altura + Idade, data = criancas)
```

```
> predict(mod_exempl2, data.frame(Idade=c(13),Altura=c(157)))
```

```
1  
49.83907
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-90.8133	36.7238	-2.473	0.02934 *
Altura	0.6643	0.2730	2.433	0.03155 *
Idade	2.7961	0.8218	3.402	0.00525 **

$$Y = -90,813 + 0,664 * (Altura) + 2,796 * (Idade)$$

Residual standard error: 4.056 on 12 degrees of freedom

Multiple R-squared: 0.826, Adjusted R-squared: 0.797

F-statistic: 28.49 on 2 and 12 DF, p-value: 2.773e-05

Regressão Linear Múltipla

Referências:

Zar, Jerrold H. Biostatistical Analysis. 4^a. Ed. New Jersey. Prentice-Hall, Inc. 1999

Wonnacott, Thomas H. & Wonnacott, Ronald J. Regression: a second course in statistics. New York. John Wiley & Sons, Inc. 1987

Exemplo:

DOI: 10.1590/1413-81232018234.03662016

1193

Excesso de peso e fatores associados em quilomboras do médio São Francisco baiano, Brasil

Overweight and associated factors in Quilombolas
from the middle San Francisco, Bahia, Brazil

TEMAS LIVRES FREE THEMES

Ricardo Franklin de Freitas Mussi ¹

Bruno Morbeck de Queiroz ¹

Edio Luiz Petróski ²

Exemplo:

Tabela 3. Associação entre excesso de peso e fatores sociodemográficos, estilo de vida e condições de saúde em adultos quilombolas. Modelos de regressão linear bruta e ajustada. Tomé Nunes, Malhada, Bahia, Brasil, 2012 (n = 112).

Variáveis	β bruto (IC95%)	p-valor*	$\beta_{ajustado}$ (IC95%)	p-valor**	R ² ajustado
Sexo					
Feminino	Referência				
Masculino	-0,289 (-3,611;-0,831)	0,002	-0,261 (-3,195;-0,817)	0,001 ^a	0,326
Situação Conjugal					
Sem companheiro	Referência				
Com companheiro	0,218 (0,262;3,187)	0,021	0,231 (0,337;3,323)	0,017 ^a	0,114
Idade em anos	0,140 (-0,10;0,680)	0,141			
Escolaridade					
Alfabetizado	Referência				
Analfabeto	0,135 (-0,641;3,984)	0,155			
Situação laboral					
Sem remuneração	Referência				
Com remuneração	0,029 (-1,287;1,760)	0,759			
Hábito etílico					
Não	Referência				
Sim	-0,128 (-2,442;0,459)	0,178			
Hábito tabagista					
Não fumante	Referência				
Fumante	-0,106 (-2,438;0,684)	0,268			
AFTL	-0,213 (-0,011;-0,001)	0,024			
AFD	-0,018 (-0,006;0,005)	0,847			
Horas de televisão	-0,035 (-0,077;0,053)	0,717			
Autoavaliação de Saúde					
Positiva	Referência				
Negativa	0,231 (0,364;3,210)	0,014	0,186 (0,088;0,177)	0,019 ^a	0,326
Glicose capilar	0,048 (-2,202;1,312)	0,617			
PAM	0,492 (0,095;0,190)	< 0,001	0,458 (0,088;0,177)	< 0,001 ^a	0,326

* Regressão linear bruta; ** Regressão linear múltipla; ^avariáveis que permaneceram no modelo final de acordo com o modelo hierárquico; AFTL: atividade física do tempo livre (horas); AFD: atividade física de deslocamento (horas); PAM: pressão arterial média (mmHg).

Mussi RFF *et al.* Excesso de peso e fatores associados em quilombolas do médio São Francisco baiano, Brasil. *Ciência & Saúde Coletiva*, 23(4):1193-1200, 2018

Exemplo:

“A Tabela 3 apresenta os resultados da análise de regressão linear bruta e múltipla. Neste sentido, a análise bruta indica que o aumento do IMC se associa ao sexo feminino, autoavaliação negativa de saúde, gastar menos tempo em atividades físicas de tempo livre e apresentar maior pressão arterial média. Durante a análise de regressão linear múltipla, foi verificada que permaneceu a associação, com à variável desfecho (IMC), o sexo feminino, a autoavaliação negativa de saúde e a pressão arterial média (PAM) aumentada. Houve uma correlação linear positiva do IMC com à pressão arterial média e a autoavaliação de saúde, e negativa com o sexo masculino. “

Regressão Logística

Regressão Logística

Muitas vezes, principalmente na área da saúde, temos interesse em identificar os fatores associados a presença ou ausência de determinada característica, como uma doença, por exemplo.

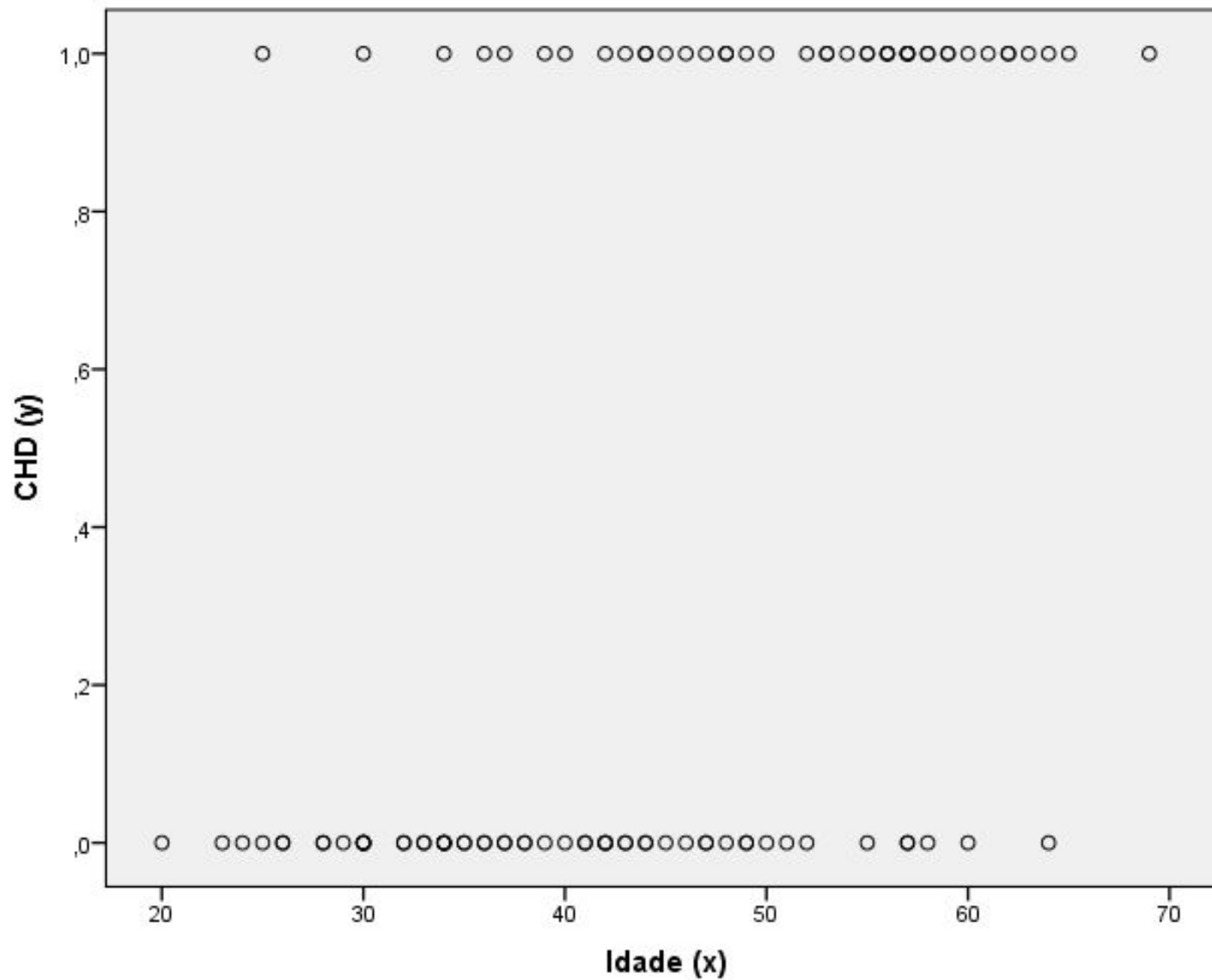
Regressão Logística

Neste caso, a variável dependente é dicotômica, assumindo o valor 0 (zero) na ausência da característica e o valor 1 (um) na sua presença.

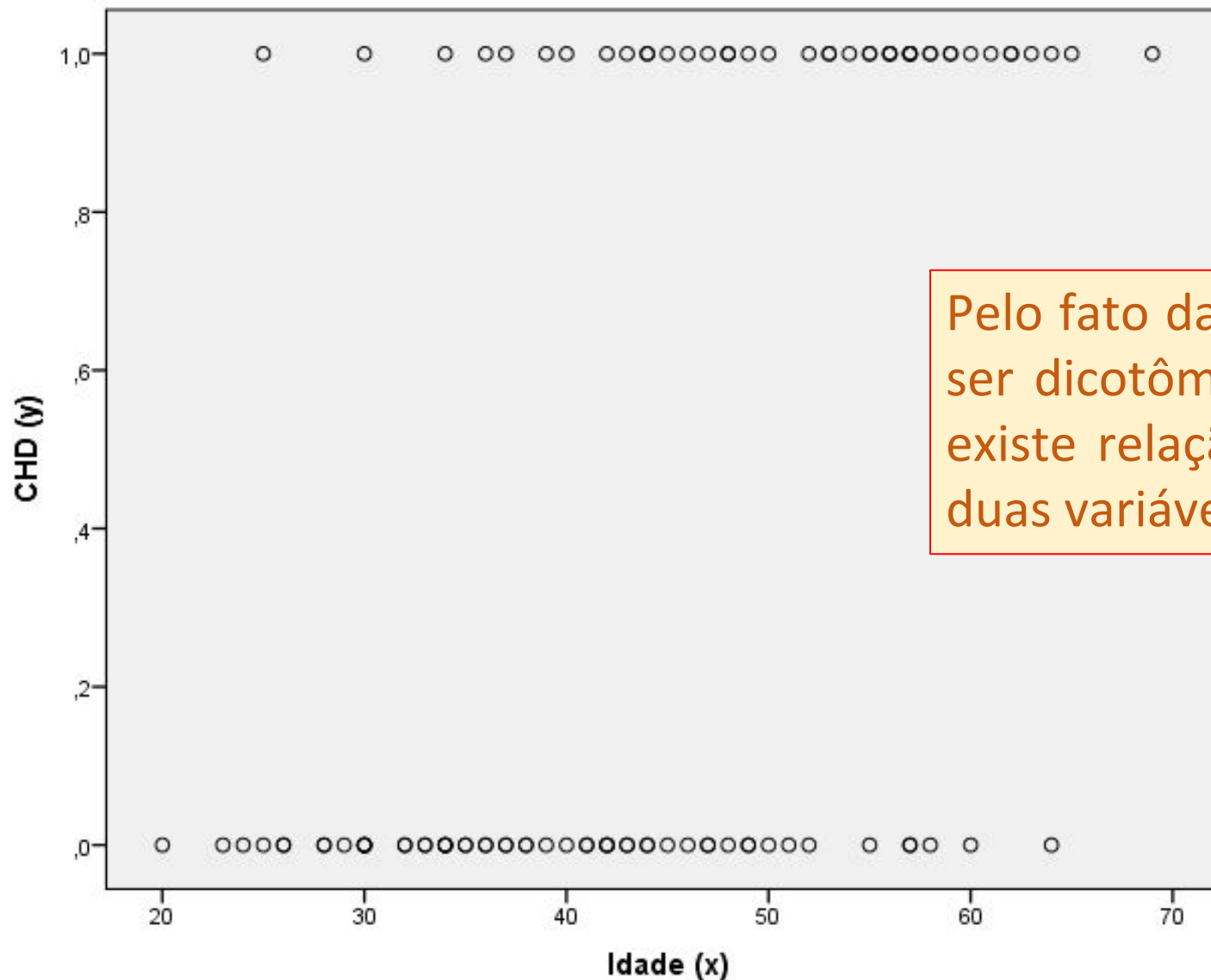
Regressão Logística

Exemplo: Usando os dados sobre presença ou ausência de doença coronária (CHD) e idade (AGE) de 100 indivíduos (Fonte: Hosmer & Lemeshow; 1989), desejamos analisar a relação entre a CHD e idade, tendo como variável dependente a CHD e como variável independente a idade.

Regressão Logística



Regressão Logística



Pelo fato da variável dependente ser dicotômica, não fica claro se existe relação funcional entre as duas variáveis.

Regressão Logística

Uma forma de visualizar melhor se existe relação entre CHD e idade é agrupar a variável dependente e analisar a média de CHD (proporção) em cada grupo de idade.

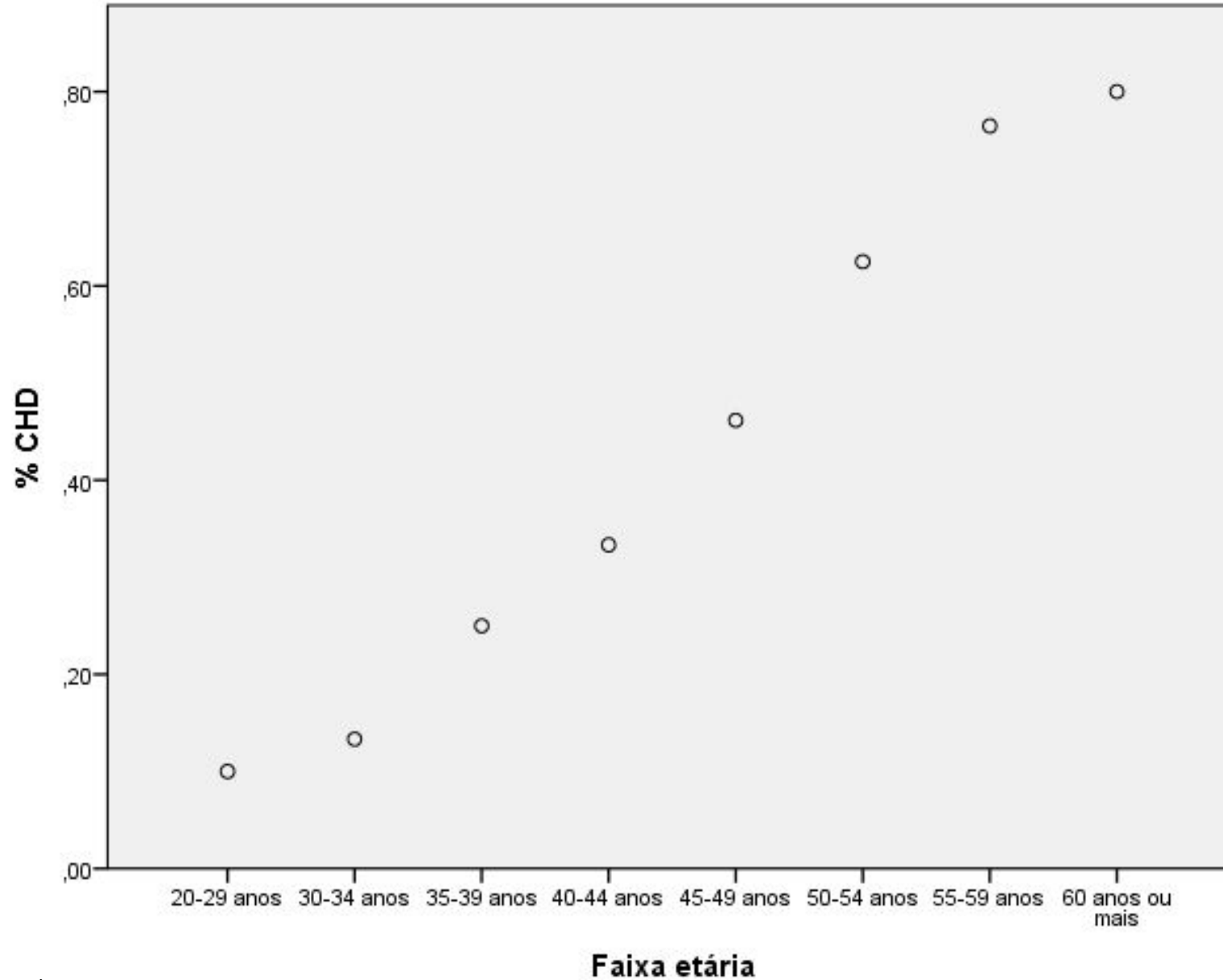
Regressão Logística

		CHD (y)					
		Não		Sim		Total	
		Contagem	N % da linha	Contagem	N % da linha	Contagem	N % da linha
Faixa etária	20-29 anos	9	90,0%	1	10,0%	10	100,0%
	30-34 anos	13	86,7%	2	13,3%	15	100,0%
	35-39 anos	9	75,0%	3	25,0%	12	100,0%
	40-44 anos	10	66,7%	5	33,3%	15	100,0%
	45-49 anos	7	53,8%	6	46,2%	13	100,0%
	50-54 anos	3	37,5%	5	62,5%	8	100,0%
	55-59 anos	4	23,5%	13	76,5%	17	100,0%
	60 anos ou mais	2	20,0%	8	80,0%	10	100,0%
Total		57	57,0%	43	43,0%	100	100,0%

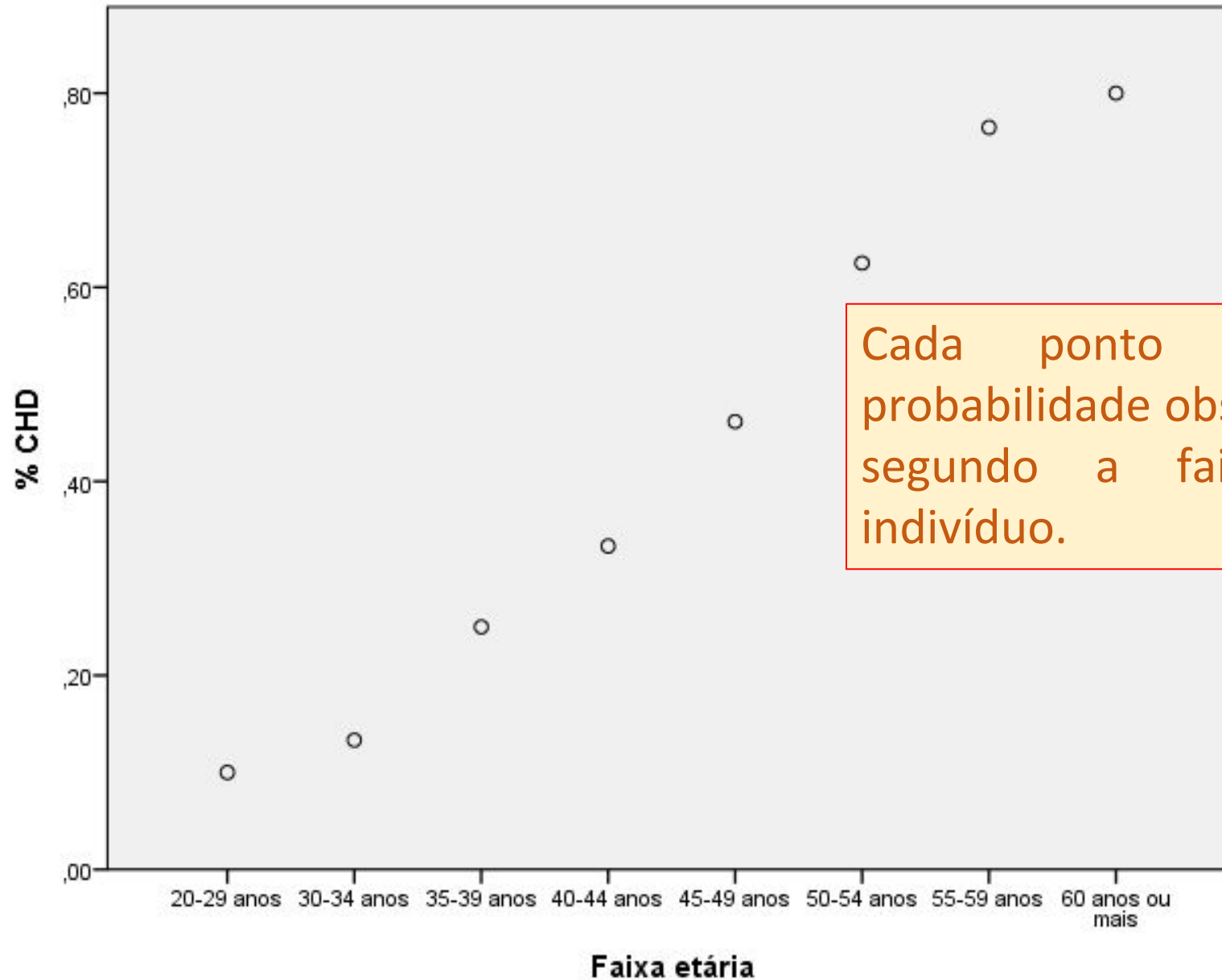
Regressão Logística

		CHD (y)					
		Não		Sim		Total	
		Contagem	N % da linha	Contagem	N % da linha	Contagem	N % da linha
Faixa etária	20-29 anos	9	90,0%	1	10,0%	10	100,0%
	30-34 anos	13	86,7%	2	13,3%	15	100,0%
	35-39 anos	9	75,0%	3	25,0%	12	100,0%
	40-44 anos	10	66,7%	5	33,3%	15	100,0%
	45-49 anos	7	53,8%	6	46,2%	13	100,0%
	50-54 anos	3	37,5%	5	62,5%	8	100,0%
	55-59 anos	4	23,5%	13	76,5%	17	100,0%
	60 anos ou mais	2	20,0%	8	80,0%	10	100,0%
Total		57	57,0%	43	43,0%	100	100,0%

Regressão Logística

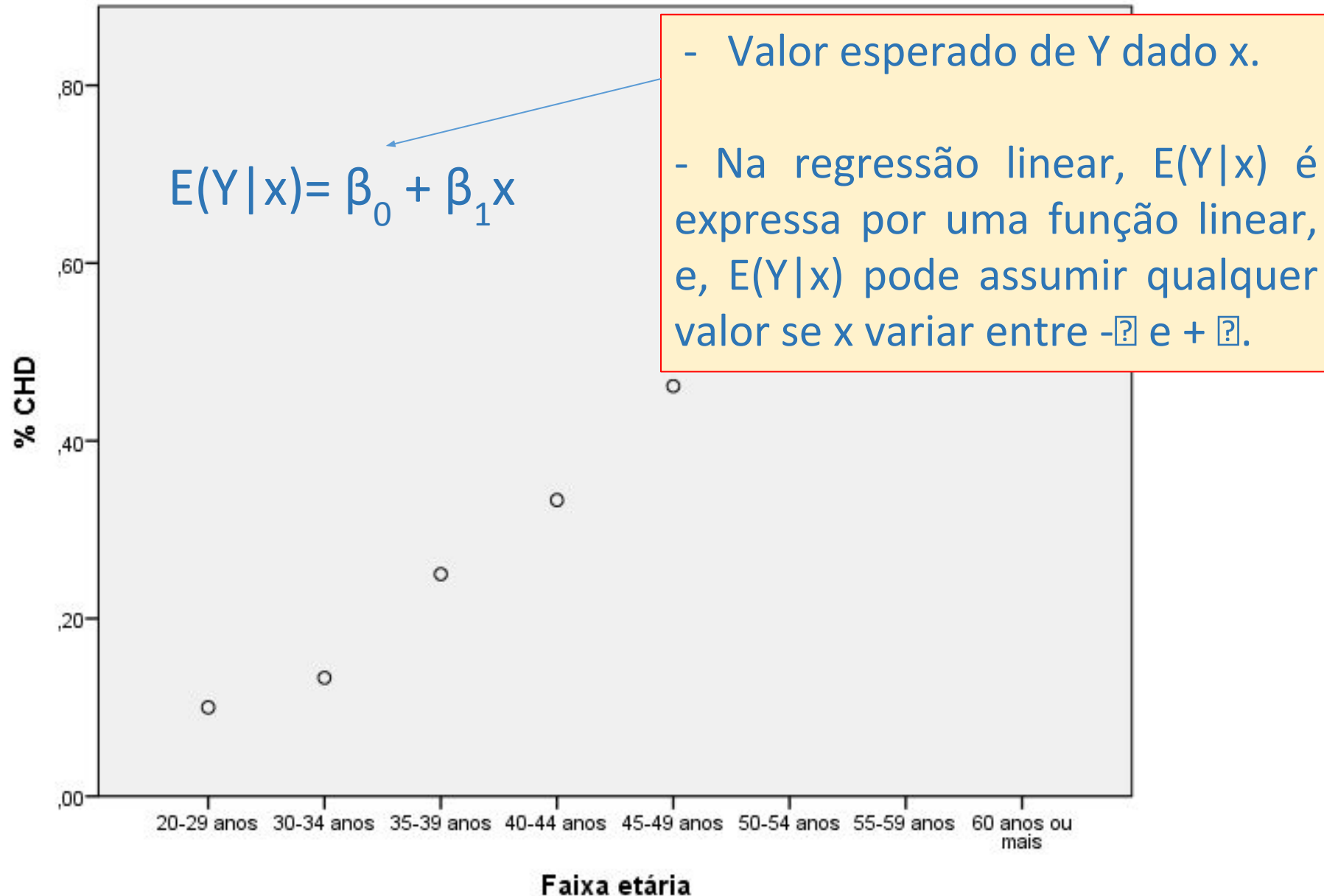


Regressão Logística

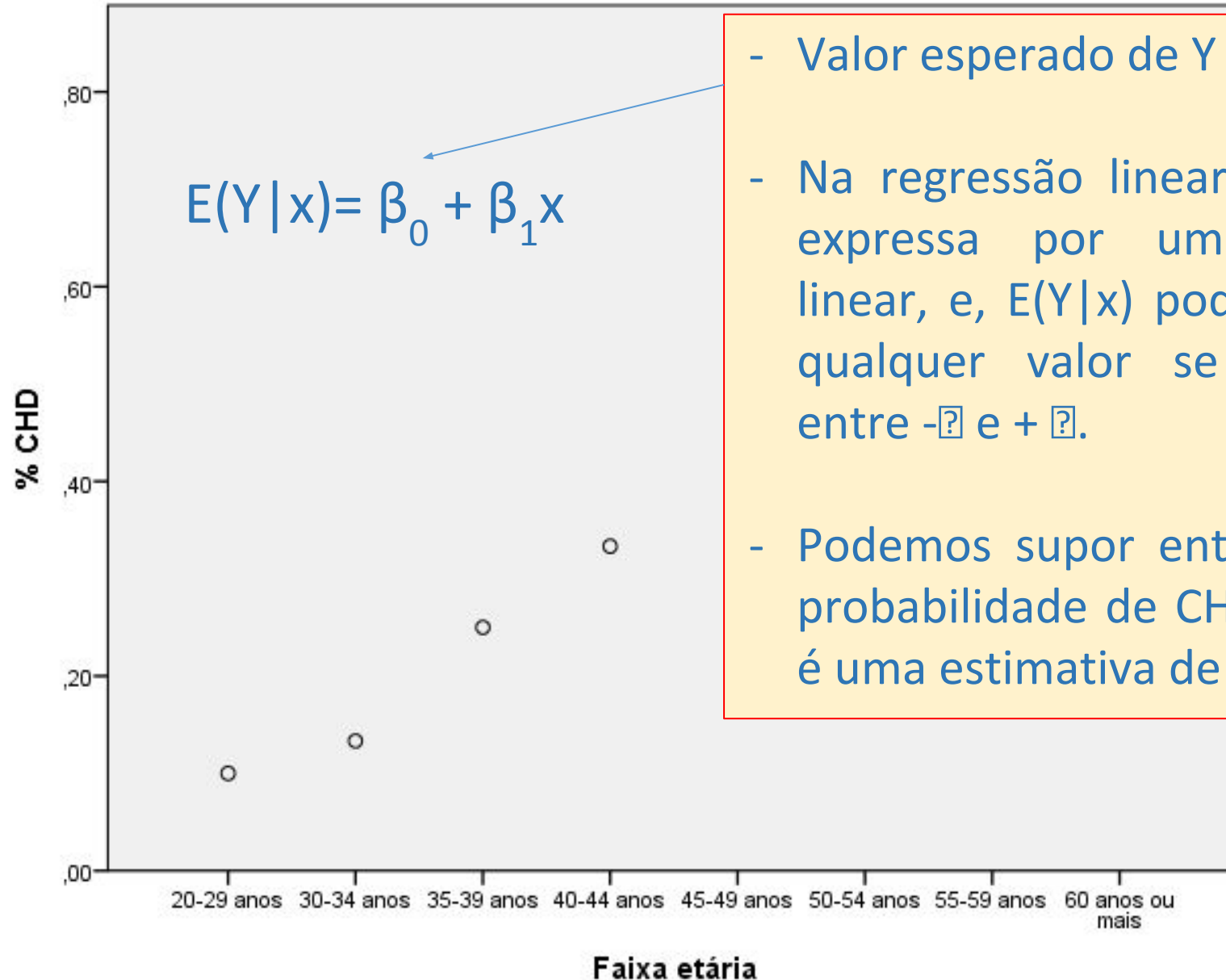


Cada ponto representa a probabilidade observada de CHD, segundo a faixa etária do indivíduo.

Regressão Logística

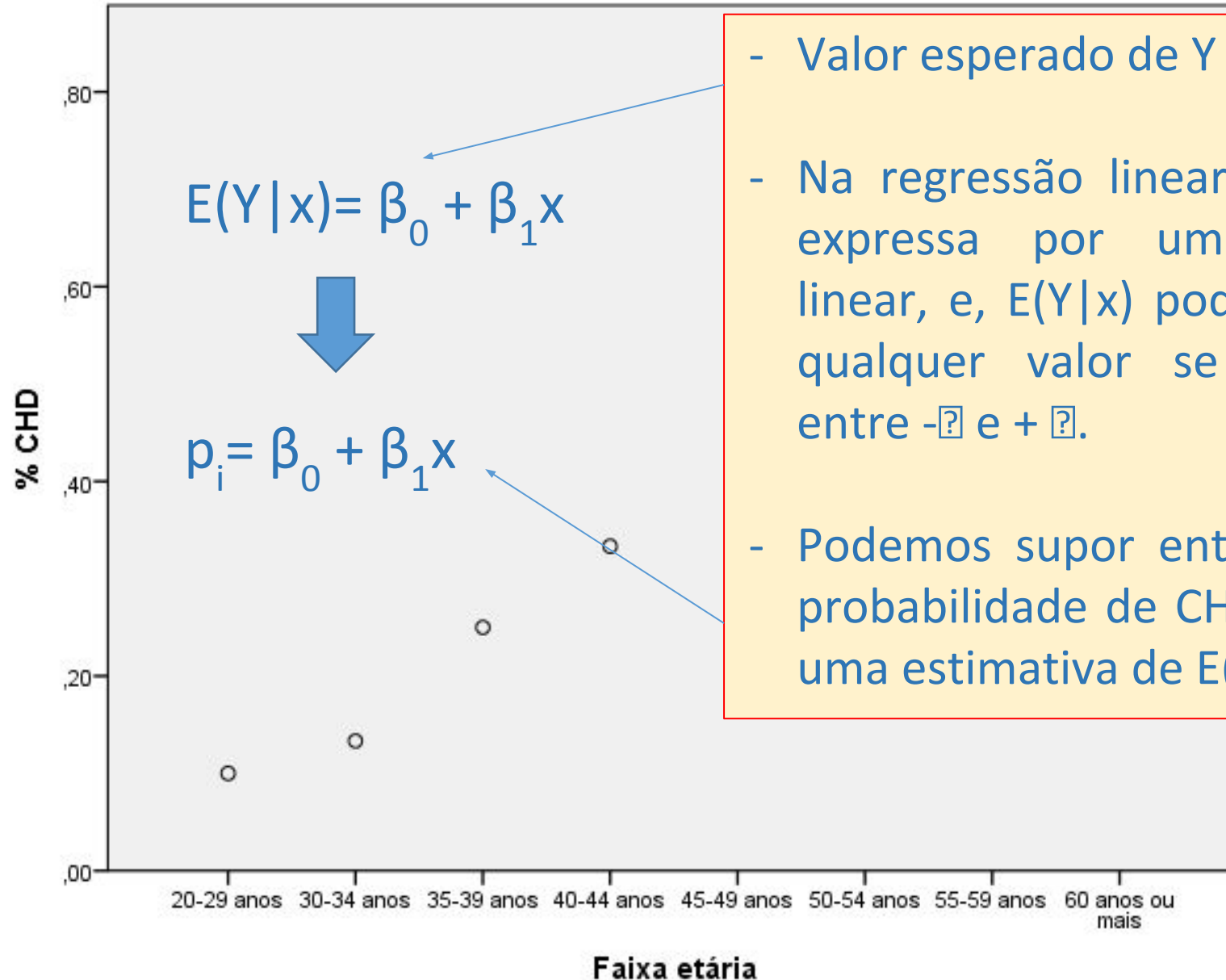


Regressão Logística



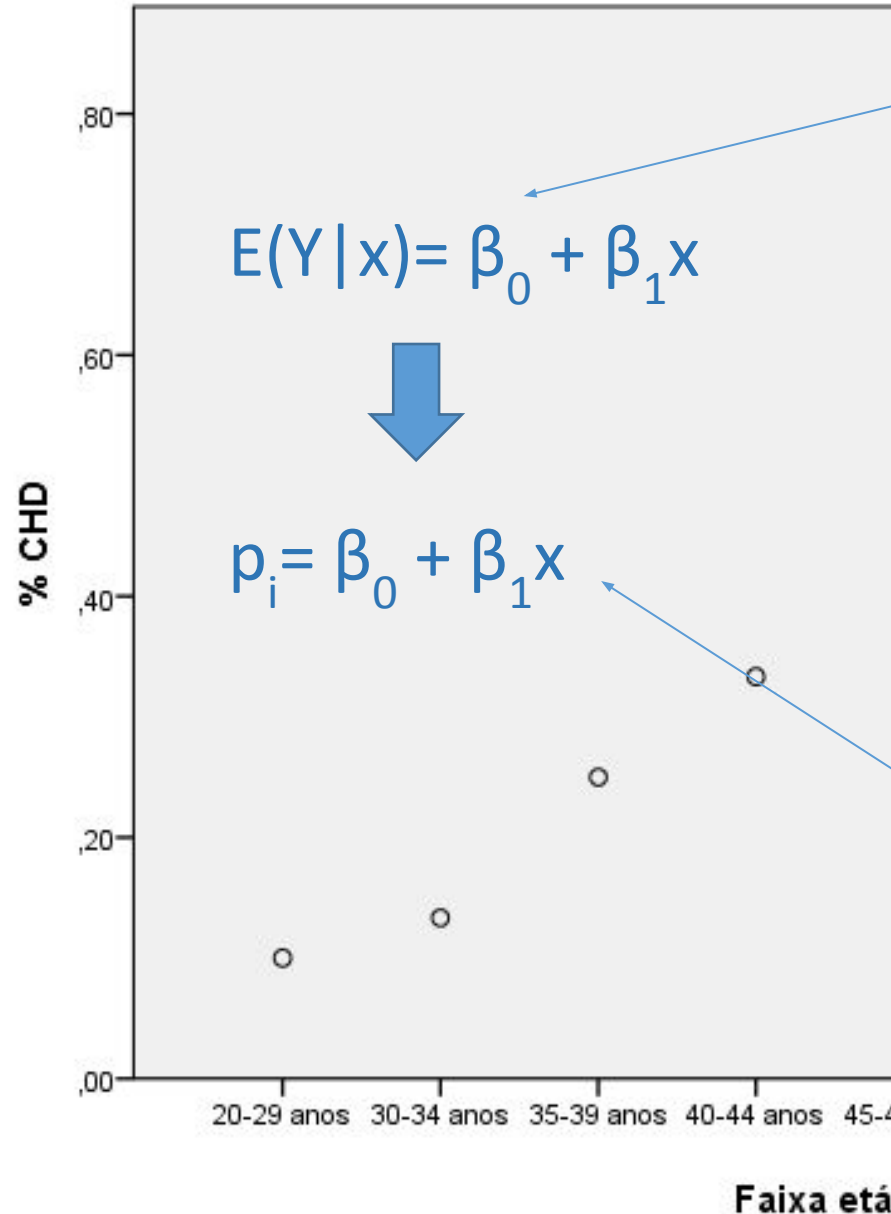
- Valor esperado de Y dado x.
- Na regressão linear, $E(Y|x)$ é expressa por uma função linear, e, $E(Y|x)$ pode assumir qualquer valor se x variar entre $-\infty$ e $+\infty$.
- Podemos supor então, que a probabilidade de CHD (eixo Y) é uma estimativa de $E(Y|x)$.

Regressão Logística



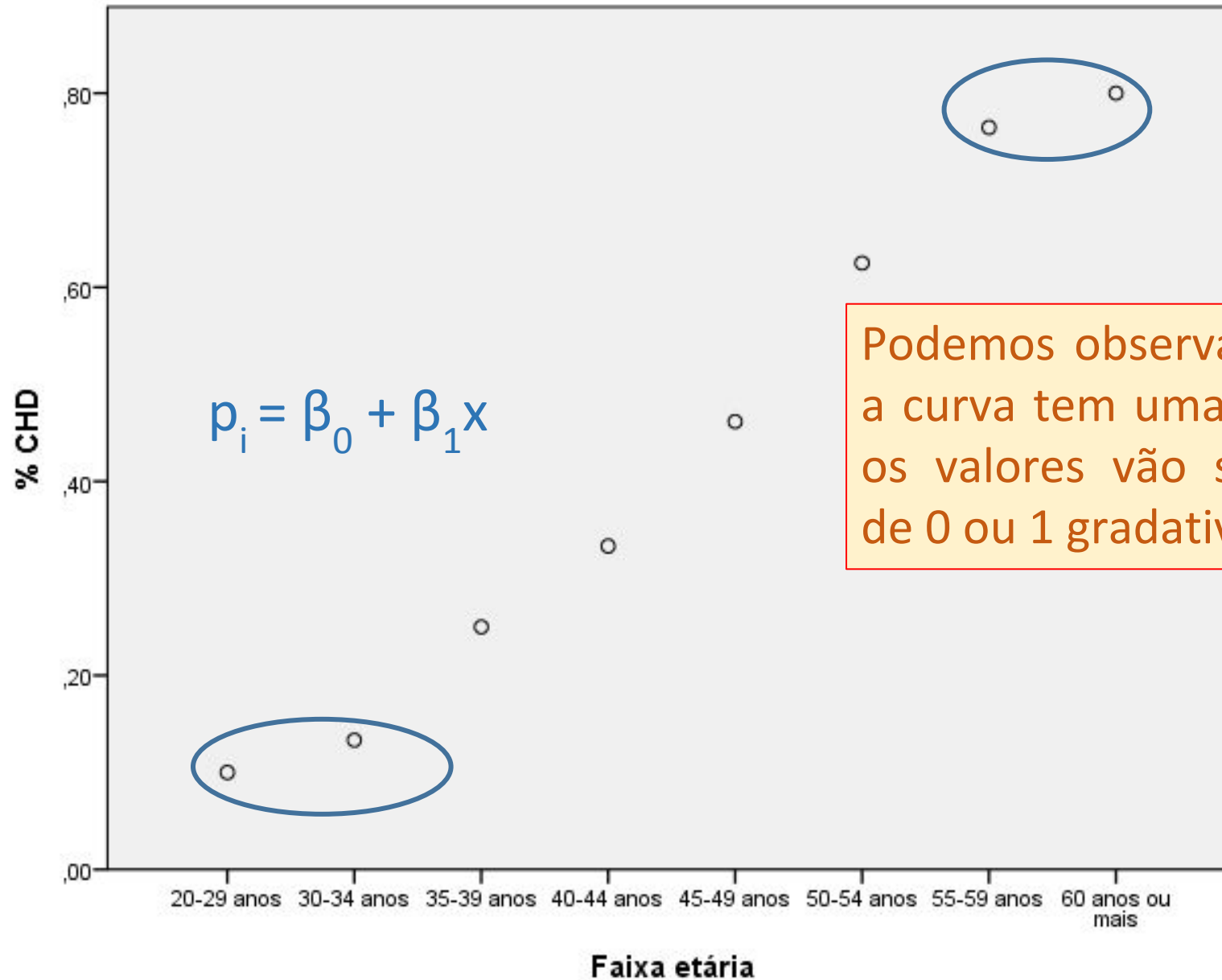
- Valor esperado de Y dado x.
- Na regressão linear, $E(Y|x)$ é expressa por uma função linear, e, $E(Y|x)$ pode assumir qualquer valor se x variar entre $-\infty$ e $+\infty$.
- Podemos supor então, que a probabilidade de CHD (eixo Y) uma estimativa de $E(Y|x)$.

Regressão Logística



- Valor esperado de Y dado x .
- Na regressão linear, $E(Y|x)$ é expressa por uma função linear, e, $E(Y|x)$ pode assumir qualquer valor se x variar entre $-\infty$ e $+\infty$.
- Podemos supor então, que a probabilidade de CHD (eixo Y) uma estimativa de $E(Y|x)$.
- Porém, neste caso, os possíveis valores de $E(Y|x)$ variam entre 0 e 1, diferente da regressão linear.

Regressão Logística



Regressão Logística

Para resolver este tipo de problema vários modelos foram criados para lidar com variável resposta dicotômica, porém, o mais utilizado é o modelo logístico.

Regressão Logística

Um dos componentes do modelo logístico é a chance de ocorrência de um evento (odds).

A chance (odds) de ocorrência de um evento pode ser definida como a razão entre o número esperado de vezes que o evento ocorrerá sobre o número esperado de vezes de que ele não ocorrerá.

Regressão Logística

Sendo assim,

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$$

Onde p_i é a probabilidade de $Y_i=1$, ou seja $\Pr(Y_i=1)$.

Regressão Logística

Sendo assim,

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$$

Onde p_i é a probabilidade de $Y_i=1$, ou seja $\Pr(Y_i=1)$.

Denominado: logito ou log-odds

Regressão Logística

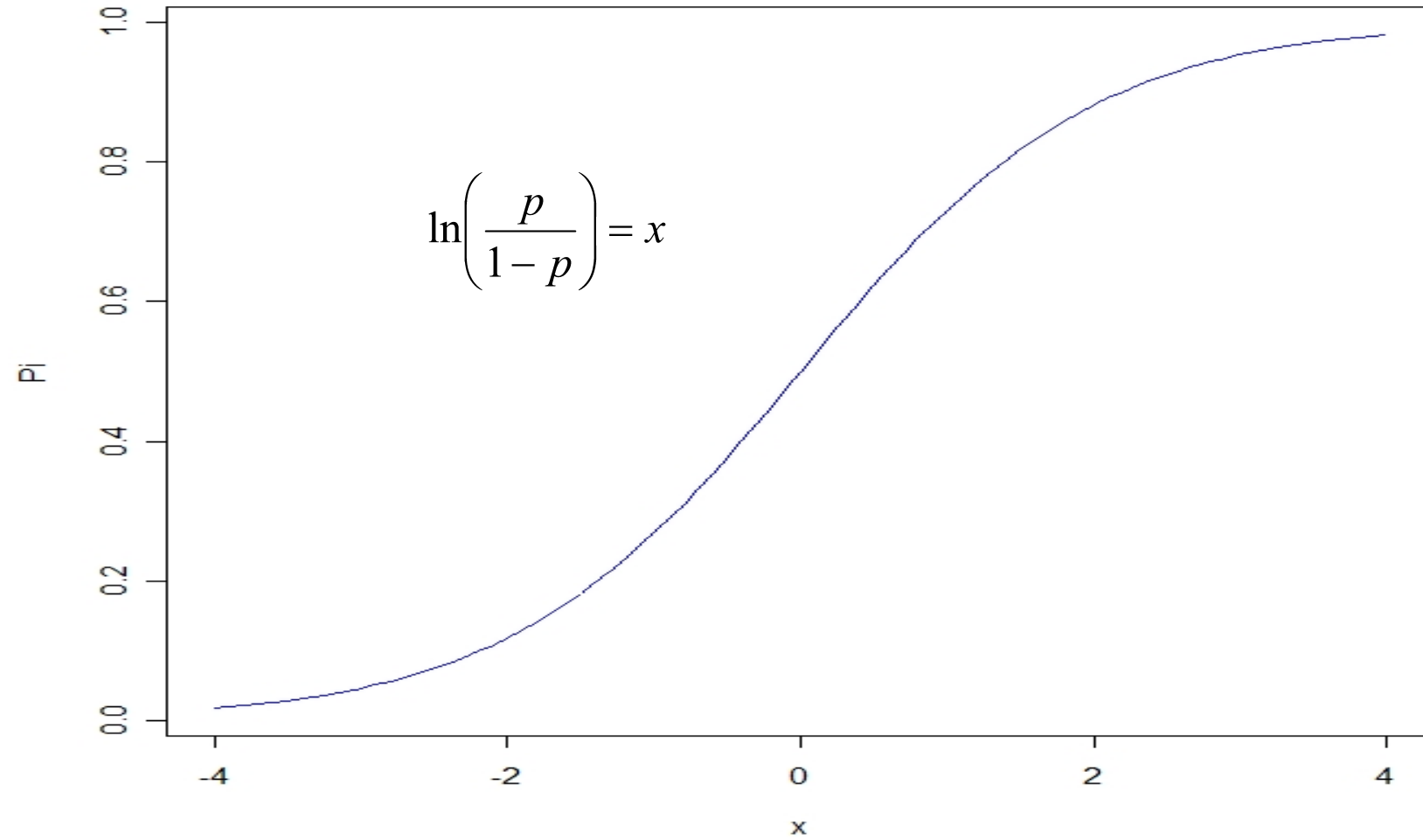
Sendo assim,

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$$

Onde p_i é a probabilidade de $Y_i=1$, ou seja $\Pr(Y_i=1)$.

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}$$

Regressão Logística



Regressão Logística

Modelo logístico com uma variável independente

$$g(x) = \ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 x$$

- $p = \Pr(Y=1)$
- Y é a variável dependente, dicotômica com valores 0 ou 1, sendo 0 a ausência da característica e 1 a presença da característica
- x é a variável independente (preditora), que pode ser quantitativa ou qualitativa

Regressão Logística

Se $x=0$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * 0 = \beta_0 \rightarrow \frac{p}{1-p} = e^{\beta_0}$$


Se $x=1$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1 \rightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1}$$

Regressão Logística


Se $x=0$

Chance do evento ($Y=1$) quando $x=0$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * 0 = \beta_0 \rightarrow \frac{p}{1-p} = e^{\beta_0}$$


Se $x=1$

Chance do evento ($Y=1$) quando $x=1$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1 \rightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1}$$



Regressão Logística

Razão de chance (OR):

$$OR = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = \frac{e^{\beta_0} * e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Regressão Logística

Razão de chance (OR):

$$OR = \frac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = \frac{e^{\beta_0} * e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$


A OR pode ser obtida pela exponencial da estimativa de β_i , ou seja, do parâmetro de x_i .

Regressão Logística

Ajustando o modelo logístico:

- Variável dependente: **CHD** -> Presença (1) ou ausência (0) de doença coronária;
- Variável independente: **fxet_55** -> Faixa etária menor 55 anos (0) ou faixa etária de 55 anos ou mais (1);

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

β_0

SE(β_0)

SE(β_1)

β_1

Regressão Logística

```
> OR1=exp(mod_CHD$coefficients)
```

```
> OR1
```

(Intercept)	fxet_55
0.4313725	8.1136364

```
> ICbeta1=confint.default(mod_CHD,level=0.95)
```

```
> ICbeta1
```

	2.5 %	97.5 %
(Intercept)	-1.340717	-0.3408489
fxet_55	1.057639	3.1294530

```
> ICOR1=exp(ICbeta1)
```

```
> ICOR1
```

	2.5 %	97.5 %
(Intercept)	0.2616579	0.7111663
fxet_55	2.8795652	22.8614705

Regressão Logística

$$OR = \text{Exp}(\beta_0)$$

```
> OR1=exp(mod_CHD$coefficients)
```

```
> OR1
```

```
(Intercept)    fxet_55  
0.4313725    8.1136364
```

```
> ICbeta1=confint.default(mod_CHD,level=0.95)
```

```
> ICbeta1
```

	2.5 %	97.5 %
(Intercept)	-1.340717	-0.3408489
fxet_55	1.057639	3.1294530

```
> ICOR1=exp(ICbeta1)
```

```
> ICOR1
```

	2.5 %	97.5 %
(Intercept)	0.2616579	0.7111663
fxet_55	2.8795652	22.8614705

Regressão Logística

```
> OR1=exp(mod_CHD$coefficients)
```

```
> OR1
```

```
(Intercept)    0.4313725  
fxet_55      8.1136364
```

```
> ICbeta1=confint.default(mod_CHD,level=0.95)
```

```
> ICbeta1
```

	2.5 %	97.5 %
(Intercept)	-1.340717	-0.3408489
fxet_55	1.057639	3.1294530

```
> ICOR1=exp(ICbeta1)
```

```
> ICOR1
```

	2.5 %	97.5 %
(Intercept)	0.2616579	0.7111663
fxet_55	2.8795652	22.8614705

$$OR = \text{Exp}(\beta_0)$$

Interpretação: um indivíduo com 55 anos ou mais possui uma chance 8 vezes maior de ter CHD quando comparado com um indivíduo com menos de 55 anos.

Regressão Logística

```
> OR1=exp(mod_CHD$coefficients)
```

```
> OR1
```

(Intercept)	fxet_55
0.4313725	8.1136364

$$OR = \text{Exp}(\beta_0)$$

```
> ICbeta1=confint.default(mod_CHD,level=0.95)
```

```
> ICbeta1
```

	2.5 %	97.5 %
(Intercept)	-1.340717	-0.3408489
fxet_55	1.057639	3.1294530

```
> ICOR1=exp(ICbeta1)
```

```
> ICOR1
```

	2.5 %	97.5 %
(Intercept)	0.2616579	0.7111663
fxet_55	2.8795652	22.8614705

IC(95%) da OR

$$\text{Exp}(\beta_1 \pm z_{1-\alpha/2} * SE(\beta_1))$$

$$\text{Exp}(2,094 \pm (1,96 * 0,529))$$

Regressão Logística

```
> OR1=exp(mod_CHD$coefficients)
```

```
> OR1
```

(Intercept)	fxet 55
0.4313725	8.1136364

```
> ICbeta1=confint.default(mod_CHD,level=0.95)
```

```
> ICbeta1
```

	2.5 %	97.5 %
(Intercept)	-1.340717	-0.3408489
fxet_55	1.057639	3.1294530

```
> ICOR1=exp(ICbeta1)
```

```
> ICOR1
```

	2.5 %	97.5 %
(Intercept)	0.2616579	0.7111663
fxet_55	2.8795652	22.8614705

$$OR = \text{Exp}(\beta_0)$$

IC(95%) da OR

Interpretação: Por ser uma razão, o valor igual a 1 representa a ausência do efeito, pois a razão entre dois valores iguais, ou seja duas chances iguais, é 1. Neste caso, deve verificar se a unidade (1) está contida no IC. No nosso exemplo, o valor 1 não pertence ao intervalo, indicando que o efeito é diferente de 1 (com 95% de probabilidade).

Regressão Logística

```
> OR1=exp(mod_CHD$coefficients)
```

```
> OR1
```

(Intercept)	fxet 55
0.4313725	8.1136364

```
> ICbeta1=confint.default(mod_CHD,level=0.95)
```

```
> ICbeta1
```

	2.5 %	97.5 %
(Intercept)	-1.340717	-0.3408489
fxet_55	1.057639	3.1294530

```
> ICOR1=exp(ICbeta1)
```

```
> ICOR1
```

	2.5 %	97.5 %
(Intercept)	0.2616579	0.7111663
fxet_55	2.8795652	22.8614705

$$OR = \text{Exp}(\beta_0)$$

IC(95%) da OR

Interpretação: Em outras palavras, na ausência do efeito (associação), a odds quando $x=0$ é igual a odds quando $x=1$, ou seja $OR=1$. Se o valor 1 está dentro do IC, significa que a OR pode assumir este valor, ou seja a associação não é significativa. Porém, o IC ao lado não contém o 1.

Regressão Logística

Teste de Wald

$$w = \frac{\beta_i}{SE(\beta_i)}$$

W tem distribuição Normal Padrão $\rightarrow N(0,1)$

*Em algumas situações o teste de Wald se comporta de maneira estranha, não rejeitando a hipótese nula quando o coeficiente é significativamente diferente de zero. Por isso, o teste de Razão de Verossimilhança é mais recomendado.

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_C  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Estatística Teste de Wald

$H_0: \beta_i = 0$

$$w = \frac{\beta_i}{SE(\beta_i)} \sim N(0,1)$$

Regressão Logística

Razão de Verossimilhança

A ideia é a mesma da regressão linear, ou seja, comparar os valores observados com os preditos pelo modelo, antes e após a inclusão da variável independente (x).

No modelo logístico esta comparação é feita através do log da verossimilhança.

Regressão Logística

Razão de Verossimilhança

Para facilitar o entendimento, vamos imaginar que os valores observados da variável resposta são os valores estimados por um modelo saturado, ou seja, que contém tantos parâmetros quanto observações (um β para cada observação).

Regressão Logística

Razão de Verossimilhança

Para comparar os dois modelos é utilizado uma expressão denominada Deviance (D):

$$D = -2\ln \left[\frac{(\textit{verossimilhança do modelo ajustado})}{(\textit{verossimilhança do modelo saturado})} \right]$$

Regressão Logística

Teste de Razão de Verossimilhança

Para testar a inclusão de uma ou mais variáveis em um modelo logístico, utilizamos a estatística G:

$$G = -2\ln \left[\frac{(\text{verossimilhança do modelo sem as variáveis})}{(\text{verossimilhança do modelo com as variáveis})} \right]$$

- G tem distribuição qui-quadrado com número de graus de liberdade dado pela diferença no número de parâmetros entre os dois modelos.

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

D do modelo só com a constante

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

D do modelo só com a constante

D do modelo após a inclusão da variável faixa etária (x)

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Score iterations: 4

$-2\ln(\text{verossimilhança do modelo sem a variável})$

D do modelo só com a constante

$-2\ln(\text{verossimilhança do modelo com } x)$

D do modelo após a inclusão da variável faixa etária (x)

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***

$-2\ln(\text{verossimilhança do modelo sem a variável})$

D do modelo só com a constante

$$G = -2\ln \left[\frac{(\text{verossimilhança do modelo sem } x)}{(\text{verossimilhança do modelo com } x)} \right] = 136,663 - 117,959 = 18,704$$

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

$-2\ln(\text{verossimilhança do modelo com } x)$

D do modelo após a inclusão da variável faixa etária (x)

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***

$$G = -2\ln \left[\frac{(\text{verossimilhança do modelo sem } x)}{(\text{verossimilhança do modelo com } x)} \right] = 136,663 - 117,959 = 18,704$$

(Dispersion parameter for $D = 18,704$ $\chi^2_{1gl} = 3,84$ $p\text{-valor} < 0,001$)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Regressão Logística

Akaike information criterion (AIC)

Assim como a Deviance, o AIC é uma medida de qualidade do ajuste que deve ser usada para comparação entre modelos.

$$AIC = -2 \log L(\hat{\theta}) + 2(p)$$

ou

$$AIC = D + 2(p)$$

Onde,

D é a Deviance do modelo $[-2\ln(\text{verossimilhança do modelo com } x)]$

P é o número de parâmetros do modelo, ou seja, número de β 's do modelo.

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

p = 2, ou seja, dois parâmetros no modelo β_0 e β_1

D

$$AIC = 117,96 + 2 (2) = 121,96$$

Regressão Logística

Probabilidade predita

Os valores preditos da regressão logística são probabilidades de ocorrência do desfecho, ou seja, é a probabilidade de $Y_i=1$ ou $\Pr(Y_i=1)$:

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}$$

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Residual deviance: 117.96 on 98 degrees of freedom

AIC: 121.96

Number of Fisher Scoring iterations: 4

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value
(Intercept)	-0.8408	0.2551	-3.296
fxet_55	2.0935	0.5285	3.961

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Number of Fisher Scoring iterations: 4

β_0

```
> p1<-prob<-predict(mod_CHD, type = "response")
> table(p1)
p1
0.301369863013931    0.7777777777777535
      73              27
```

β_1

$$p_i = Pr(Y_1 = 1 | X_1 = 1) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \frac{\exp(-0,8408 + (2,0935 * 1))}{1 + \exp(-0,8408 + (2,0935 * 1))} = 0,778$$

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value
(Intercept)	-0.8408	0.2551	-3.296
fxet_55	2.0935	0.5285	3.961

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

Number of Fisher Scoring iterations: 4

β_0

```
> p1<-prob<-predict(mod_CHD, type = "response")  
> table(p1)
```

p1
0.301369863013931
0.7777777777777535
73
27

β_1

$$p_i = Pr(Y_1 = 1|X_1 = 1) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \frac{\exp(-0,8408 + (2,0935 * 1))}{1 + \exp(-0,8408 + (2,0935 * 1))} = 0,778$$

Regressão Logística

```
> mod_CHD=glm(CHD~fxet_55,data=dados_CHD, family=binomial(link="logit"))  
> summary(mod_CHD)
```

Call: glm(formula = CHD ~ fxet_55, family = binomial(link = "logit"), data = dados_CHD)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7344	-0.8469	-0.8469	0.7090	1.5488

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.8408	0.2551	-3.296	0.00098 ***
fxet_55	2.0935	0.5285	3.961	7.46e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

$$p_i = Pr(Y_1 = 1 | X_1 = 1) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \frac{\exp(-0,8408 + (2,0935 * 1))}{1 + \exp(-0,8408 + (2,0935 * 1))} = 0,778$$

Number of Fisher Scoring iterations: 4

Exemplo:

DOI: 10.1590/1980-5497201500060004

ARTIGO ORIGINAL / ORIGINAL ARTICLE

Determinantes da autoavaliação de saúde no Brasil e a influência dos comportamentos saudáveis: resultados da Pesquisa Nacional de Saúde, 2013

Determinants of self-rated health and the influence of healthy behaviors: results from the National Health Survey, 2013

Celia Landmann Szwarcwald^I, Giseli Nogueira Damacena^I, Paulo Roberto Borges de Souza Júnior^I, Wanessa da Silva de Almeida^I, Lilandra Torquato Medrado de Lima^I, Deborah Carvalho Malta^{II}, Sheila Rizzato Stopa^{II}, Maria Lúcia França Pontes Vieira^{III}, Cimar Azeredo Pereira^{III}

Exemplo:

Tabela 3. Resultados dos modelos univariado e multivariado de regressão logística tendo como desfecho a autoavaliação de saúde muito ruim/ruim. Pesquisa Nacional de Saúde, Brasil, 2013.

Variáveis	OR bruta (IC95%)	OR ajustada (IC95%)
Sexo		
Masculino	1,00	1,00
Feminino	1,44 (1,28 – 1,61)*	1,23 (1,09 – 1,38)**
Idade	1,04 (1,04 – 1,04)*	1,01 (1,01 – 1,02)*
Grau de escolaridade [#]		
1	8,98 (6,78 – 11,88)*	6,39 (4,77 – 8,55)*
2	2,50 (1,82 – 3,42)*	2,60 (1,89 – 3,58)*
3	1,68 (1,23 – 2,29)**	1,95 (1,43 – 2,66)*
4	1,00	1,00
Cor ou raça		
Branca	0,68 (0,60 – 0,76)*	0,70 (0,61 – 0,80)*
Não branca	1,00	1,00
Pelo menos uma DCNT		
Sim	7,56 (6,38 – 8,96)*	5,34 (4,48 – 6,36)*
Não	1,00	1,00

*Valor p < 1%; **Valor p < 5%.

[#]1-Sem instrução/fundamental incompleto; 2-Fundamental completo/médio incompleto; 3-Médio completo/superior incompleto; 4-Superior completo e mais. OR: razão de chances; IC95%: intervalo de confiança de 95%; DCNT: doença crônica não transmissível.

SZWARCWALD, Celia Landmann et al.
Determinantes da autoavaliação de saúde no Brasil
e a influência dos comportamentos saudáveis:
resultados da Pesquisa Nacional de Saúde, 2013.
Rev. bras. epidemiol. [online]. 2015, vol.18, suppl.2
[cited 2018-05-16], pp.33-44.

Exemplo:

“Os resultados dos modelos de regressão logística apresentados na Tabela 3, tendo como variável resposta a autoavaliação ruim/muito ruim, mostram, primeiramente, que todos os fatores sociodemográficos considerados no estudo têm efeitos significativos ($p < 0,01$). No que se refere à idade, foi evidenciada uma associação direta, isto é, quanto mais velho é o indivíduo, maior é o percentual de percepção ruim da própria saúde. Quanto às diferenças por sexo, as mulheres têm pior AAS do que os homens, e em relação à raça/cor, os indivíduos não brancos avaliam pior a sua saúde do que os brancos. Os efeitos do grau de instrução foram altamente significativos. A razão de chances (OR) de ter uma avaliação ruim/muito ruim da própria saúde foi 9 vezes maior entre os que têm ensino fundamental incompleto, quando comparados aos que completaram o ensino superior, e 7 vezes maior no modelo ajustado por idade, sexo, raça/cor e presença de pelo menos uma DCNT.

Os resultados apresentados na Tabela 3 mostram, adicionalmente, os efeitos significativos ($p < 0,01$) da presença de DCNT sobre a AAS ruim/muito ruim. A OR foi 5,3 vezes maior entre os indivíduos que tiveram diagnóstico de pelo menos uma DCNT, quando comparados aos demais, mesmo após o controle dos fatores sociodemográficos.”

Regressão Logística

Referências:

HOSMER, DW Jr. & LEMESHOW, S. Applied Logistic Regression.
1989. John Willey & Sons, Inc.