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 $y, x_1, x_2, \ldots, x_k$.

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:

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

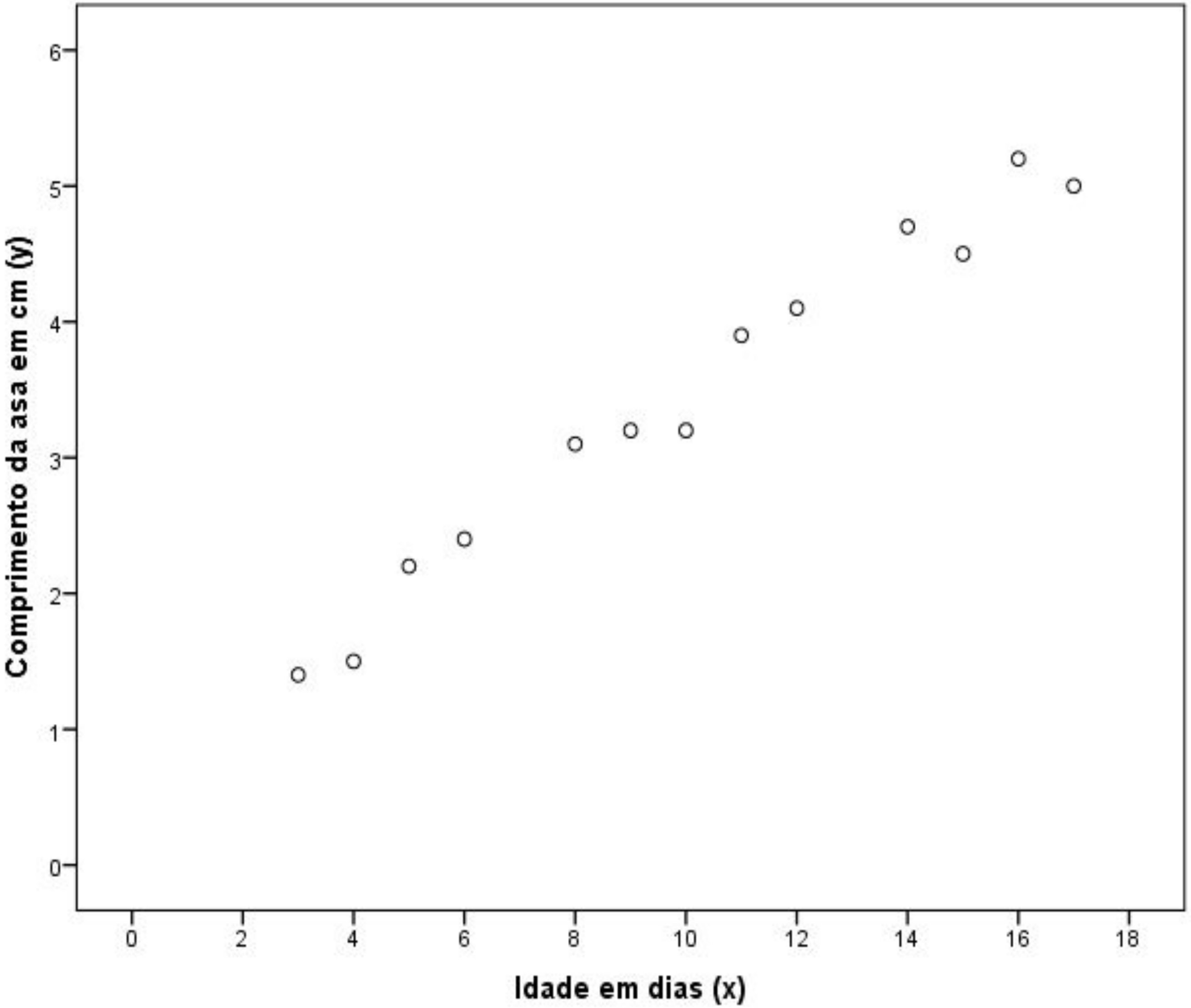
- β_0 e β_1 são constantes não conhecidas que serão estimadas a partir dos dados disponíveis
- Permite prever a variável y a partir das observações de x

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:

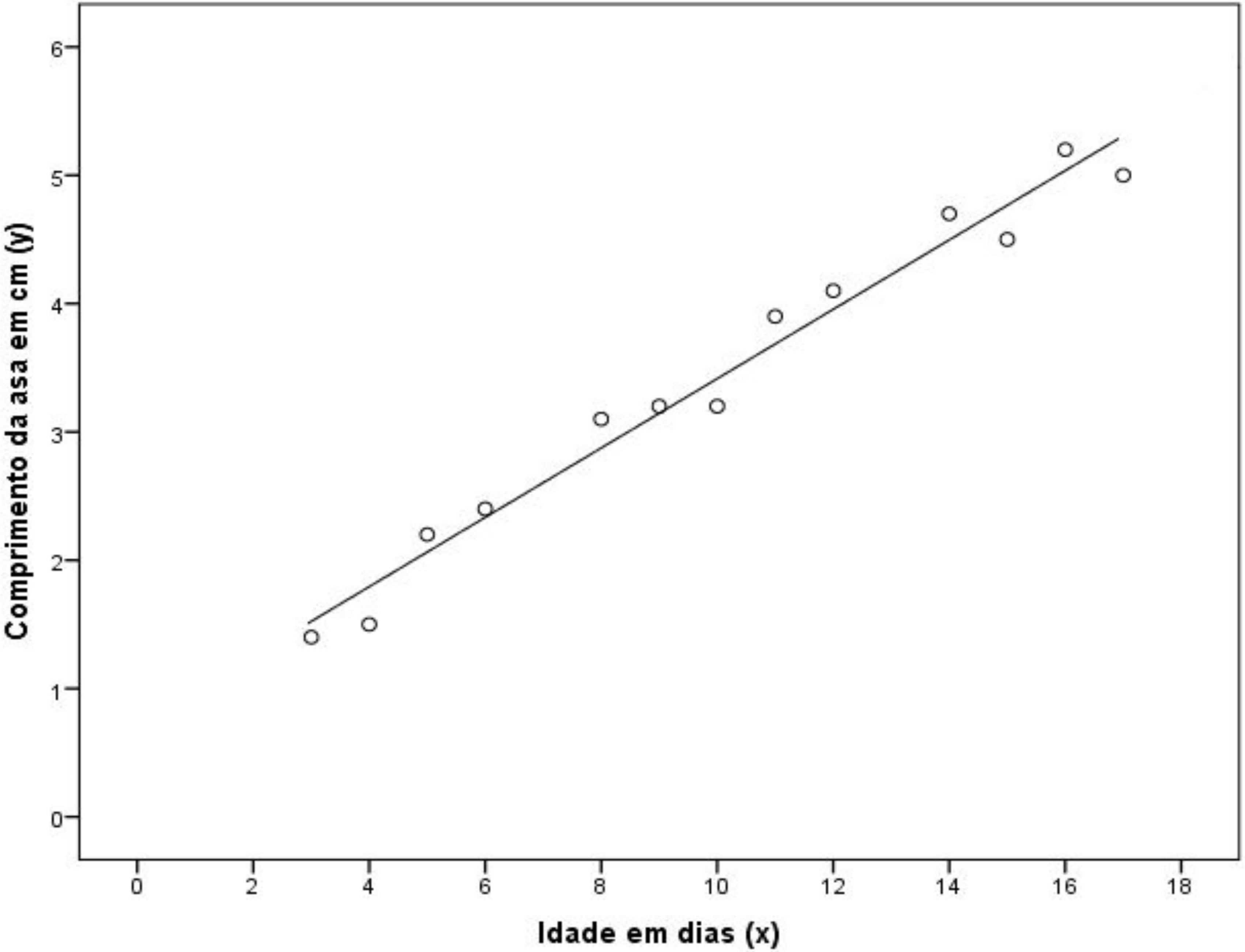
Regressão Linear Simples



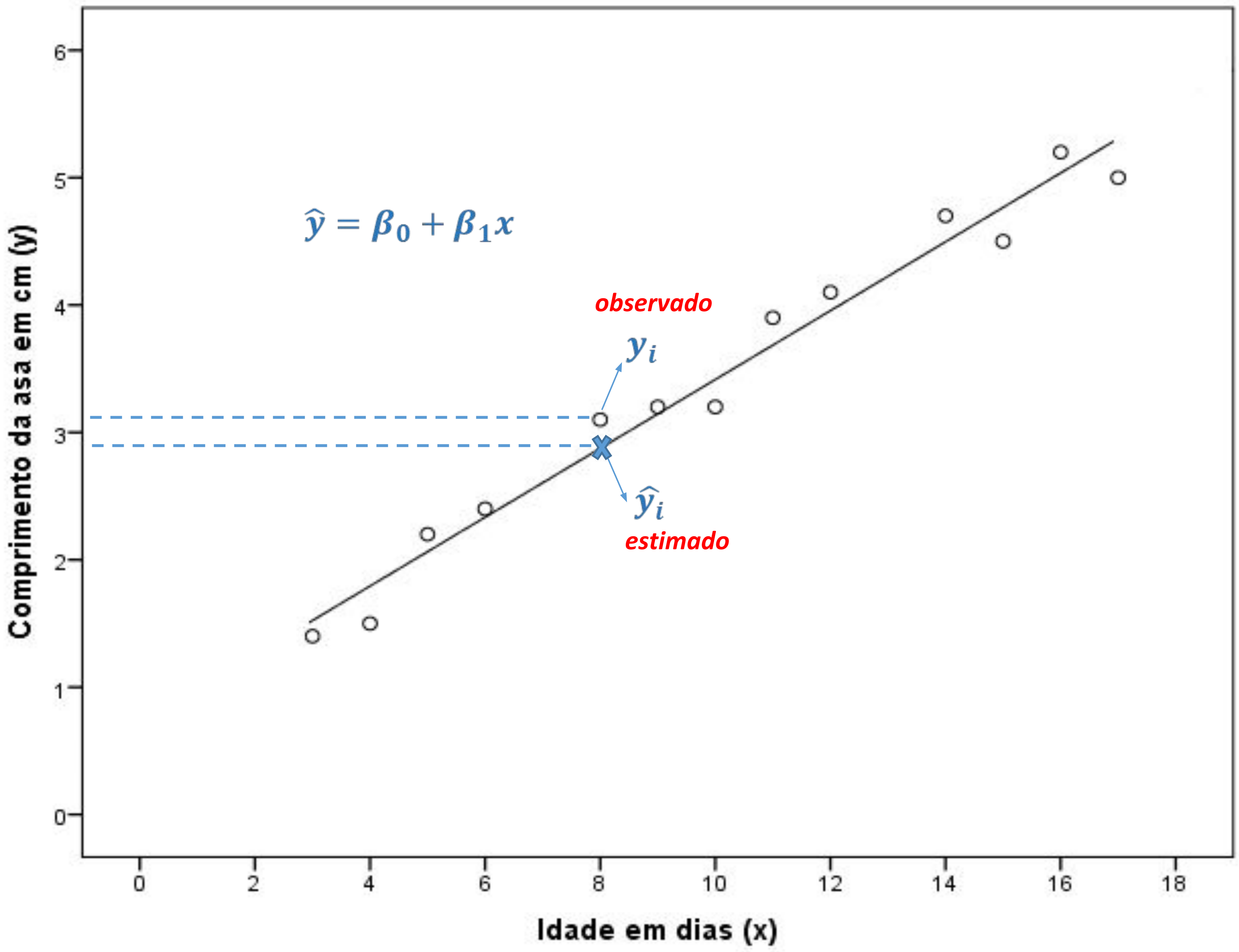
Regressão Linear Simples

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),..., (x_{13}, y_{13}).

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



Comprimento da asa em cm (y)
0
1
2
3
4
5
6
0
2
4
6
8
10
12
14
16
18
Idade em dias (x)



$\hat{y} = \beta_0 + \beta_1 x$
observado
y_i
$\hat{y}_i$
estimado
Comprimento da asa em cm (y)
Idade em dias (x)
0
1
2
3
4
5
6
0
2
4
6
8
10
12
14
16
18

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.

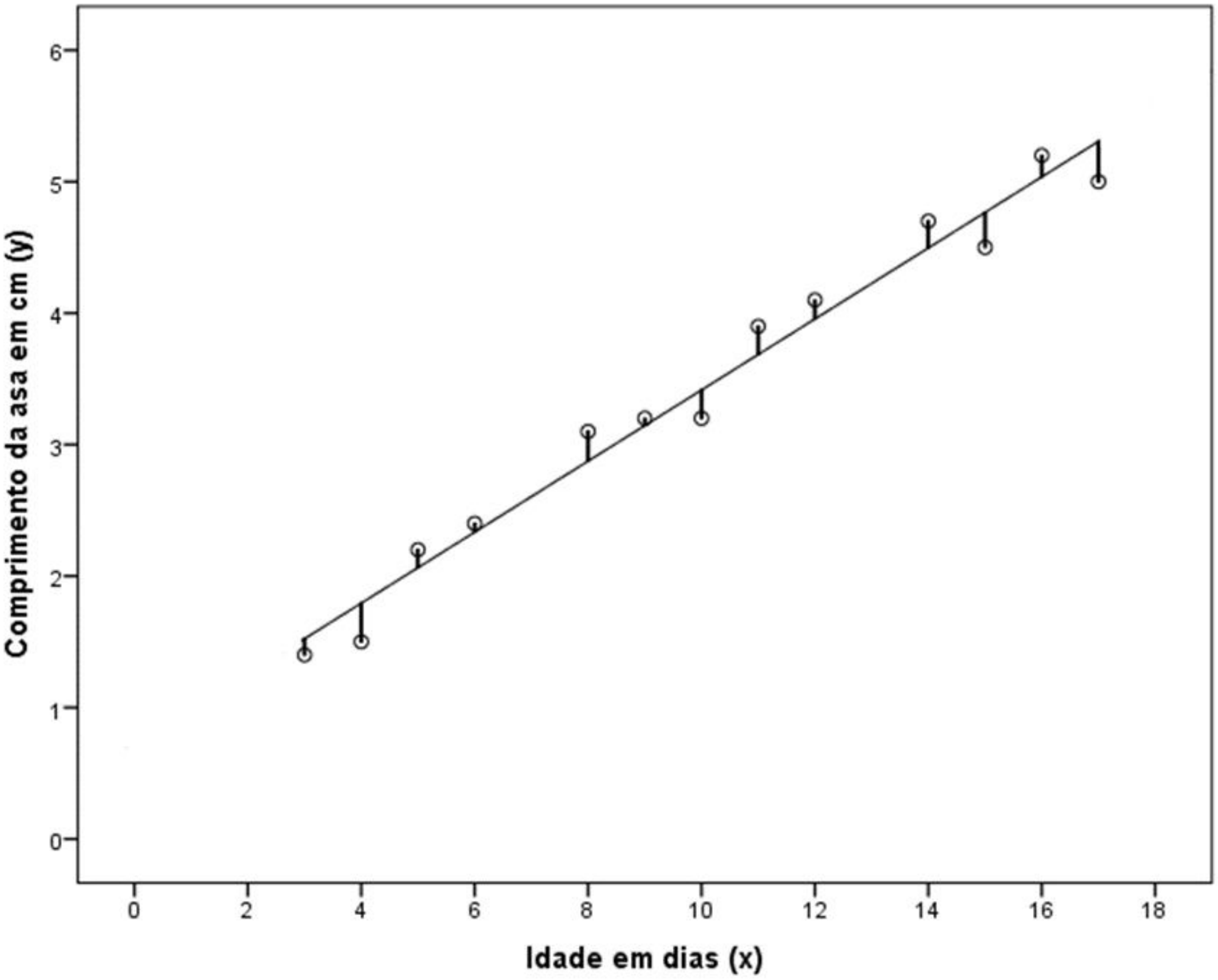
Porém, em uma nuvem de pontos mais realista, é possível traçar várias retas diferentes, cujos coeficientes β_0 e β_1 da equação são estimados por métodos distintos.

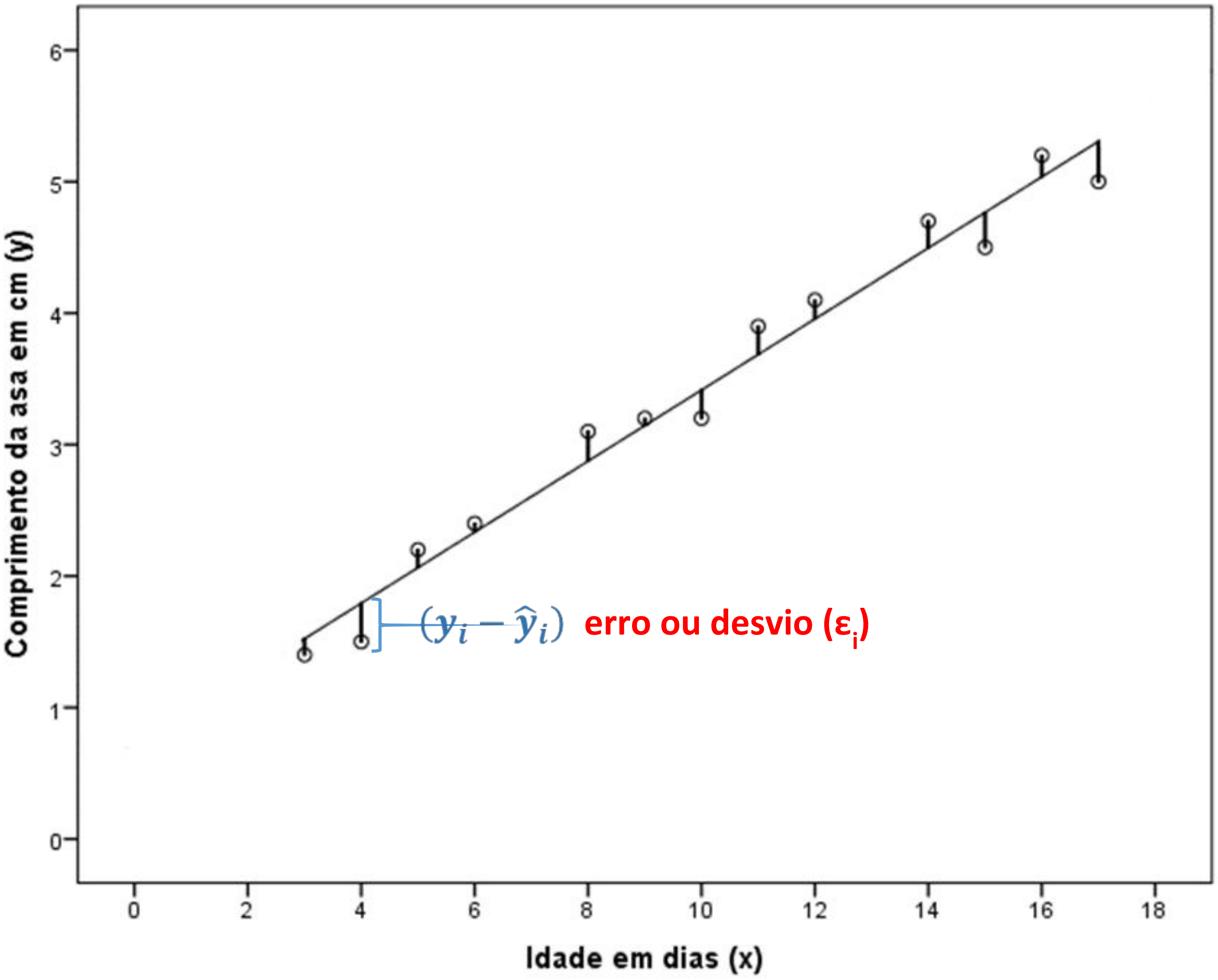
Regressão Linear Simples

- 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 $\sum_{i=1}^{n}(y_1 - \hat{y}_i)^2$ é mínimo.


Comprimento da asa em cm (y)
Idade em dias (x)
0
1
2
3
4
5
6
0
2
4
6
8
10
12
14
16
18



Comprimento da asa em cm (y)
Idade em dias (x)
0
1
2
3
4
5
6
0
2
4
6
8
10
12
14
16
18
$(y_i - \hat{y}_i)$
erro ou desvio (ε_i)

Regressão Linear Simples

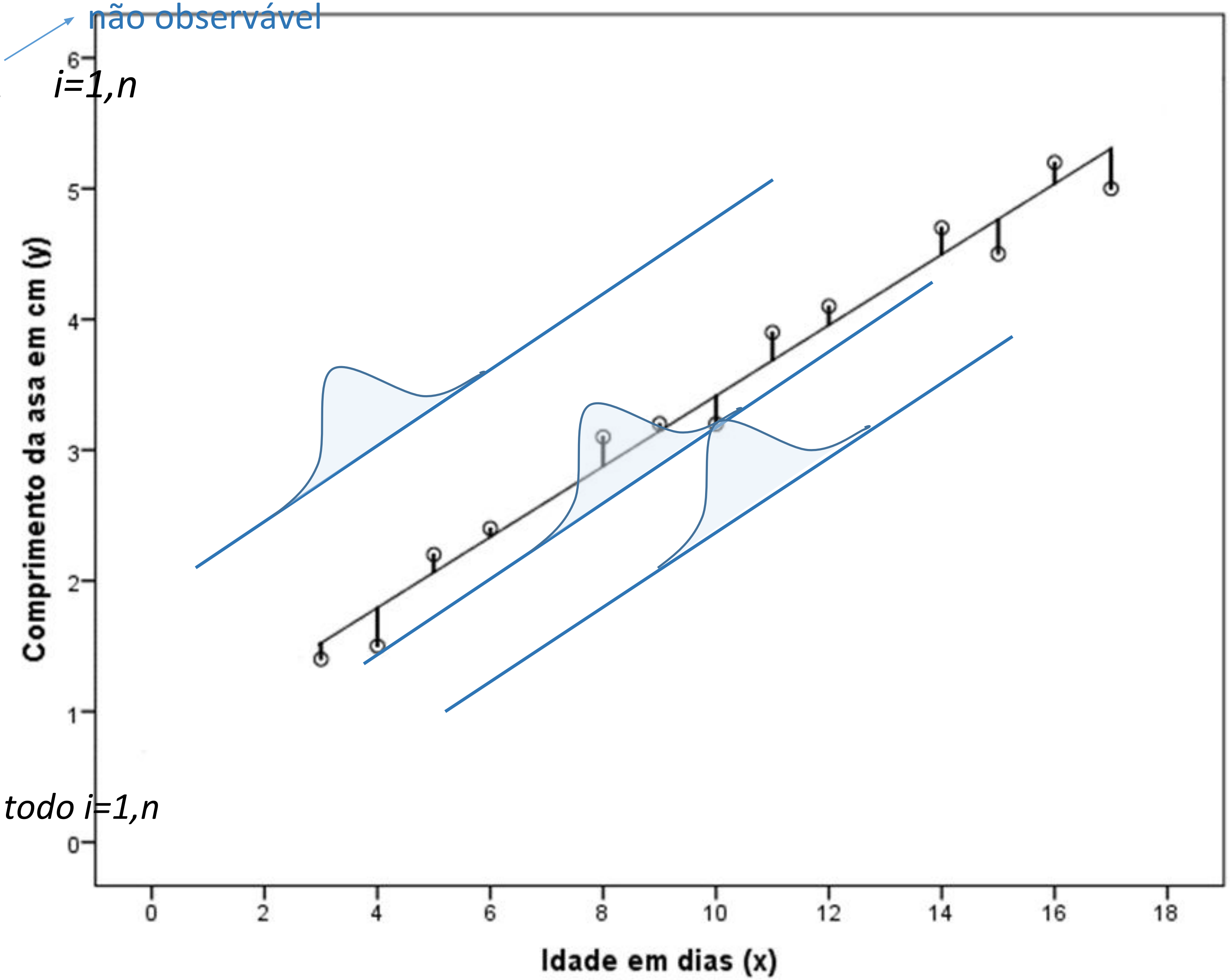
O erro **ε** pode representar:

- 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,

$$y = \beta 0 + \beta 1x \quad \text{ou} \quad y = \beta 0 + \beta 1x + \beta 1x^2?$$

Erro, variável aleatória não observável

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{Componente determinístico}} + \varepsilon_i \qquad i=1,n$$

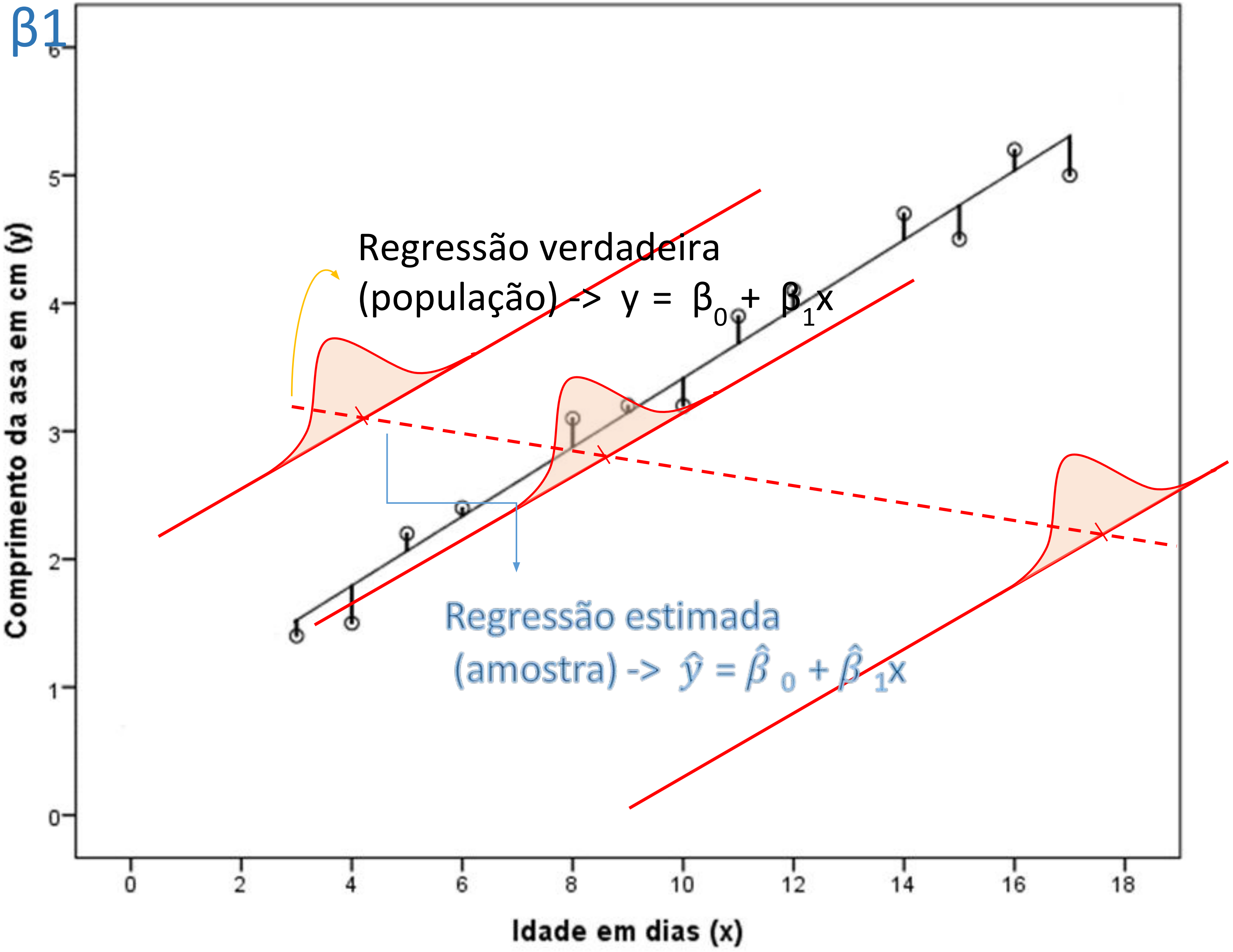


Como $\varepsilon_i \sim N(0,\sigma^2)$ *para todo i=1,n*

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Regressão Linear Simples

Estimando β0 e β1



Regressão Linear Simples

Estimando β0 e β1

Estimador de mínimos quadrados

Soma do quadrado dos erros: $f = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ como $\hat{y}_i = \beta_0 + \beta_1 x_i$ Temos:

$$f = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

As estimativas de β_0 e β_1 devem minimizar a soma dos quadrados dos desvios

$$\min_{\beta_0, \beta_1} f = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

No ponto de mínimo as derivadas parciais são nulas

$$\begin{cases} \dfrac{\partial f}{\partial \beta_0} = 0 \rightarrow \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0 \\ \dfrac{\partial f}{\partial \beta_1} = 0 \rightarrow \sum_{i=1}^{n} x_i [y_i - (\beta_0 + \beta_1 x_i)] = 0 \end{cases}$$

Regressão Linear Simples

Estimando β0 e β1

Estimador de mínimos quadrados

Sistema de equações normais
A solução deste sistema fornece os estimadores de β_0 e β_1

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Estimadores de mínimo os quadrados:

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

ou

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

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

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 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \ldots, n$;

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 comprimento 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, $Cov(\varepsilon_i,\varepsilon_k) = 0$ para todo $i \neq k$.

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

$V(\varepsilon_i) = \sigma^2$ 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.;

$\varepsilon_i \sim N(0,\sigma^2)$ 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

Estimando β0 e β1

Se as hipóteses H1 até H6 forem satisfeitas, os estimadores de mínimos quadrados são estimadores lineares não tendenciosos de variância mínima (Teorema de Gauss Markov)

Estimadores são normalmente distribuídos: $\hat{\beta}_0 \sim n(\beta_0, \sigma^2_{\beta_0}$ **e** $\hat{\beta}_1 \sim n(\beta_1, \sigma^2_{\beta_1}$

Estimadores de variâncias: $\sigma^2_{\beta_0} = \dfrac{\sigma^2_\varepsilon \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}$ **e** $\sigma^2_{\beta_1} = \dfrac{\sigma^2_\varepsilon}{\sum_{i=1}^{n} x_i^2 - \bar{x}2}$

$$\sigma^2_\varepsilon = \frac{\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2}{n - 2}$$

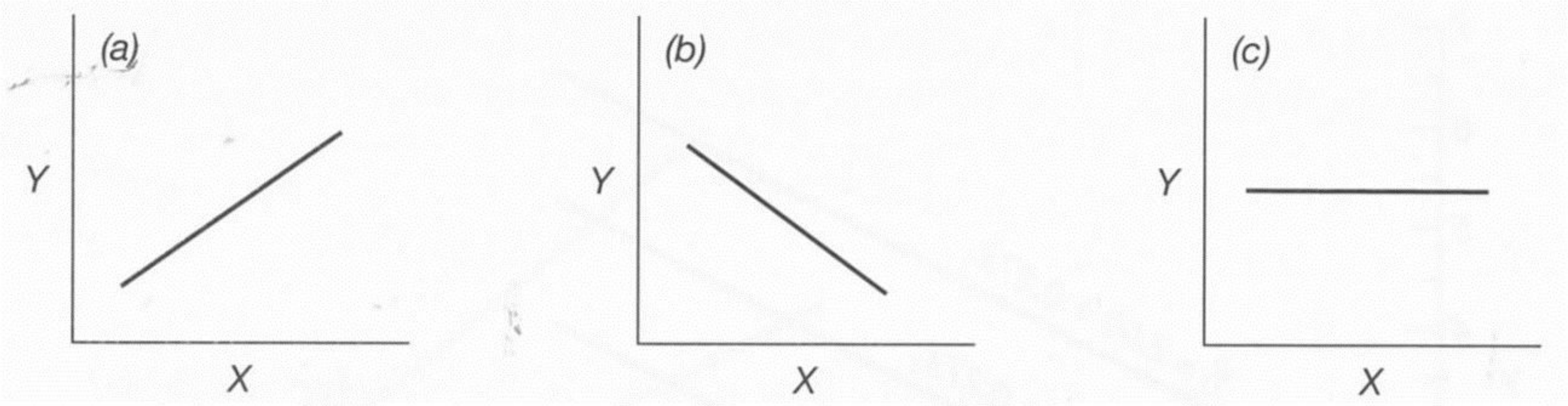
Regressão Linear Simples

Os coeficientes β0 e β1

O parâmetro β1 é o coeficiente angular da reta de regressão e define a inclinação da reta.

O valor de β1 pode ser positivo ou negativo, variando, teoricamente entre

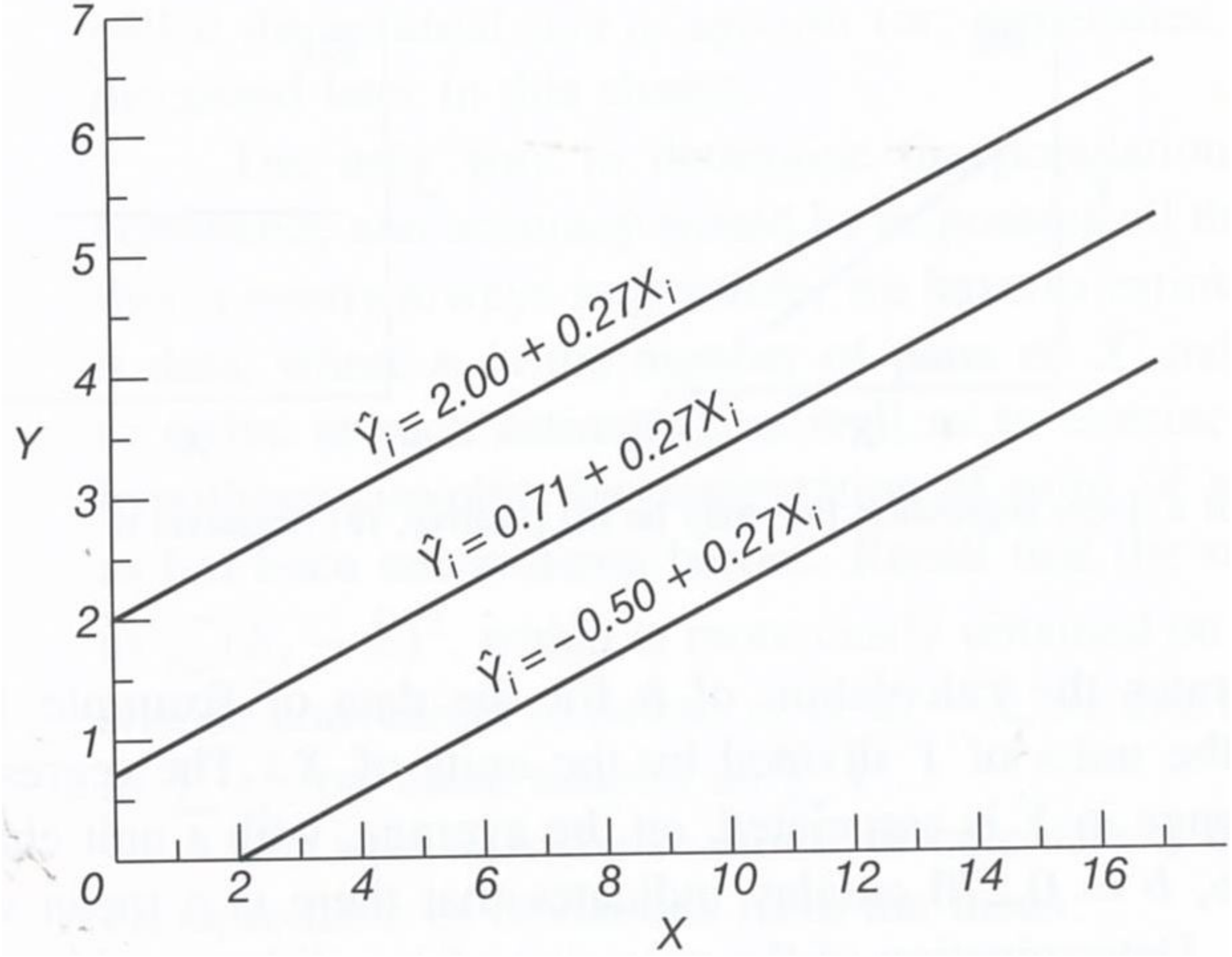
-∞ e +∞, incluindo o zero.



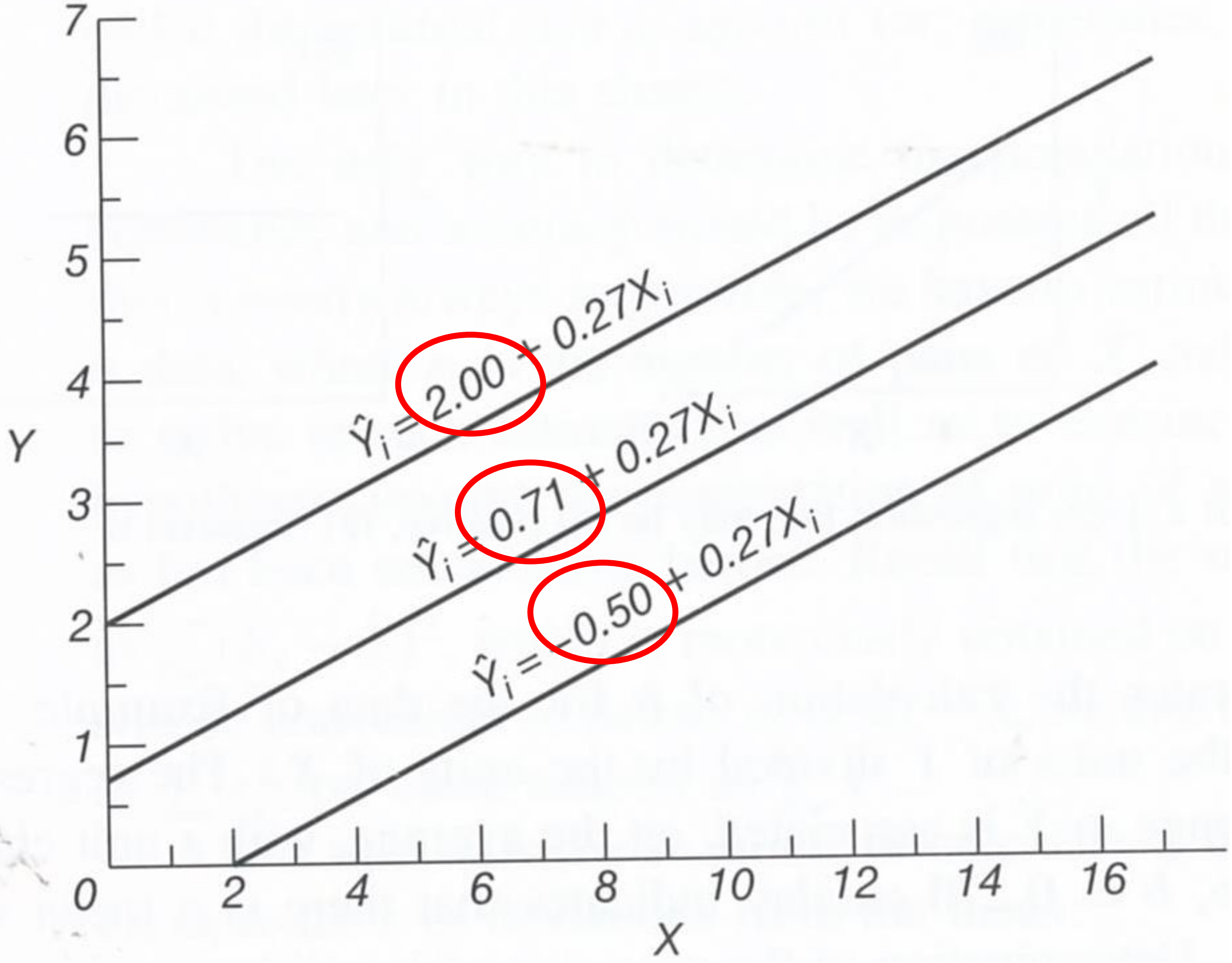
β_1 positivo β_1 negativo β_1 zero

Regressão Linear Simples

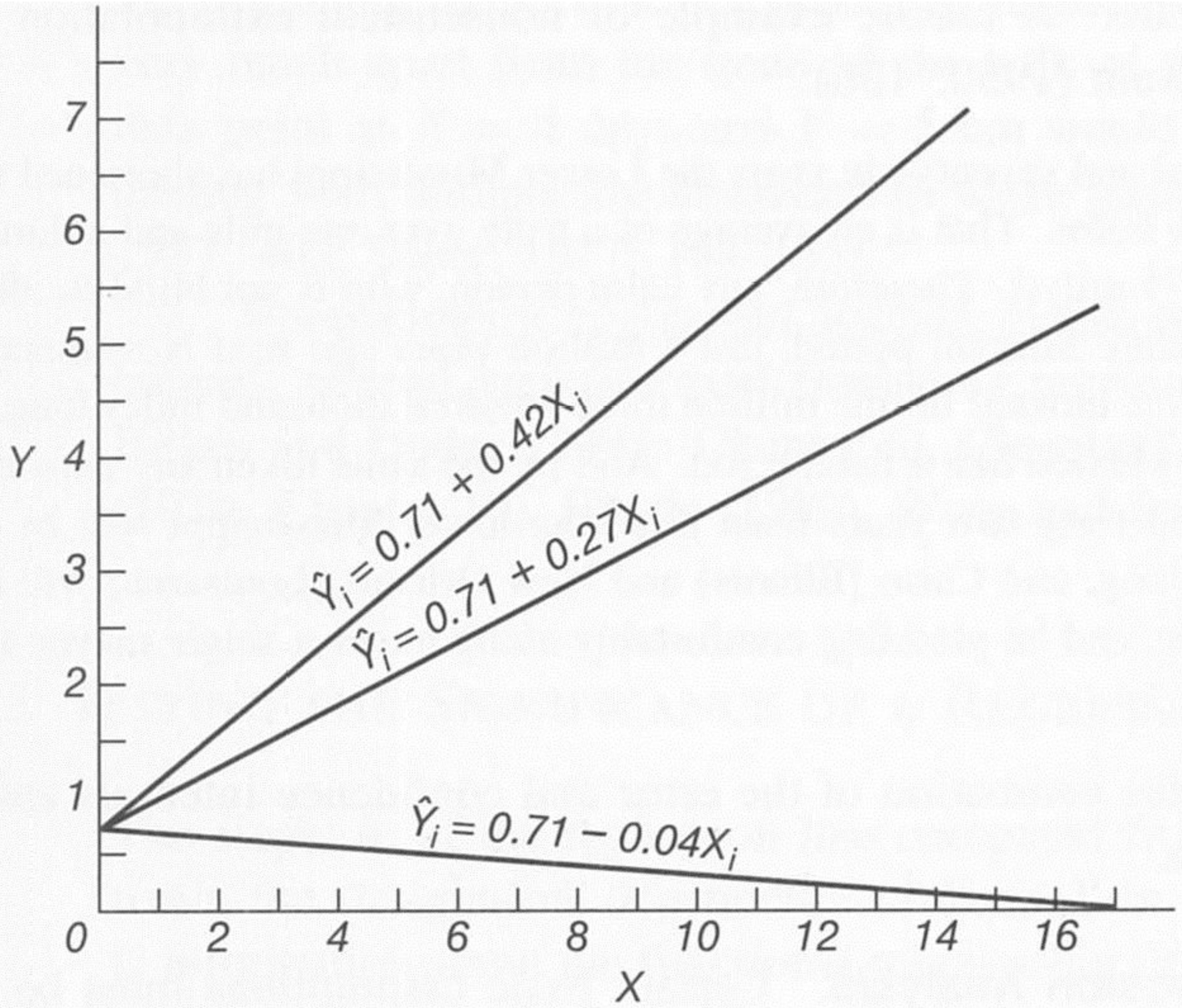
- Os coeficientes $\beta 0$ e $\beta 1$



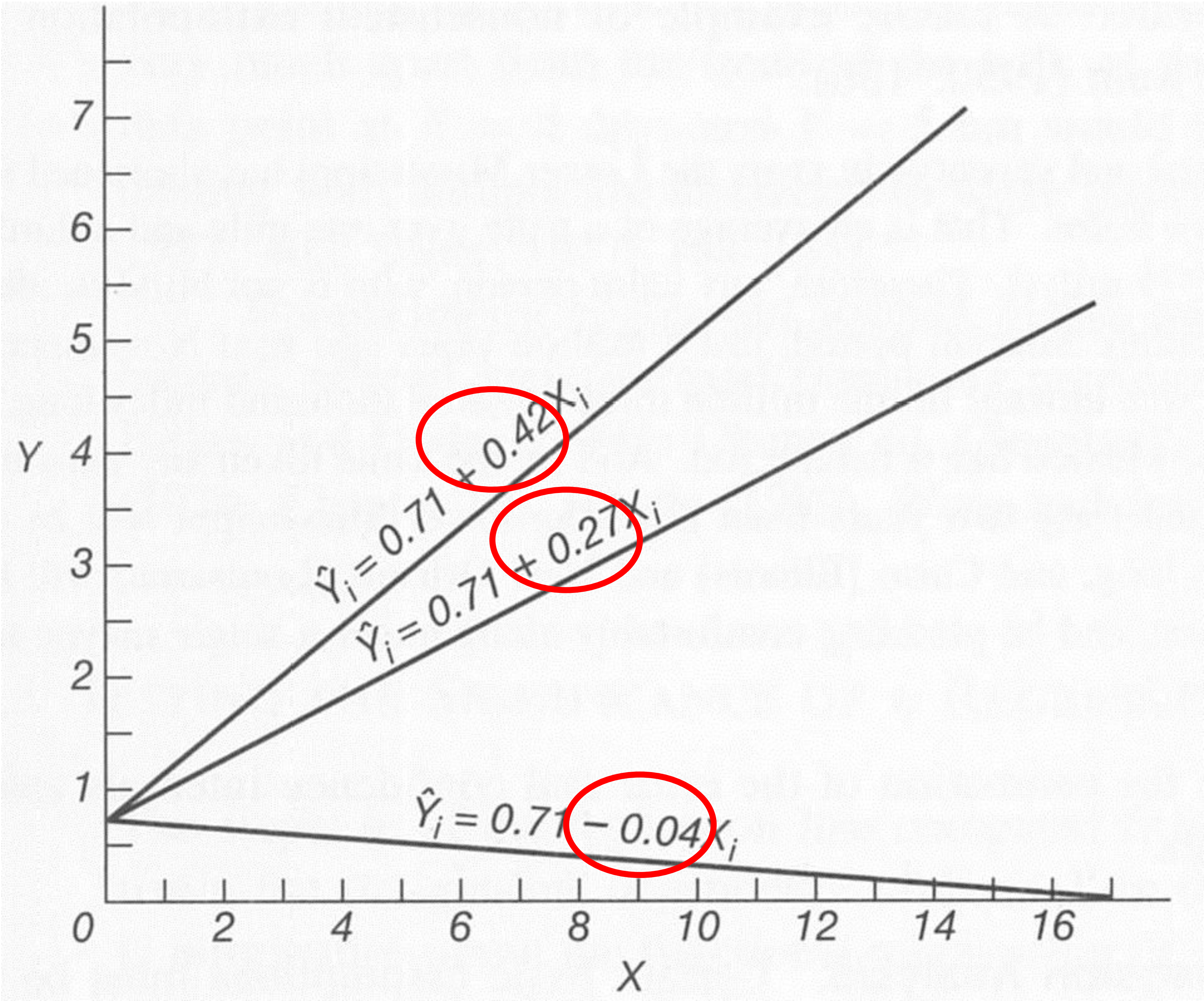
Y
X
7
6
5
4
3
2
1
0
2
4
6
8
10
12
14
16
Ŷi = 2.00 + 0.27Xi
Ŷi = 0.71 + 0.27Xi
Ŷi = −0.50 + 0.27Xi



7
6
5
4
3
2
1
0
Y
2
4
6
8
10
12
14
16
X
$\hat{Y}_i = 2.00 + 0.27X_i$
$\hat{Y}_i = 0.71 + 0.27X_i$
$\hat{Y}_i = -0.50 + 0.27X_i$



7
6
5
4
3
2
1
0
Y
2
4
6
8
10
12
14
16
X
$\hat{Y}_i = 0.71 + 0.42X_i$
$\hat{Y}_i = 0.71 + 0.27X_i$
$\hat{Y}_i = 0.71 - 0.04X_i$



Y
X
0
1
2
3
4
5
6
7
2
4
6
8
10
12
14
16
$\hat{Y}_i = 0.71 + 0.42X_i$
$\hat{Y}_i = 0.71 + 0.27X_i$
$\hat{Y}_i = 0.71 - 0.04X_i$

Regressão Linear Simples

- Calculando os valores preditos de Y

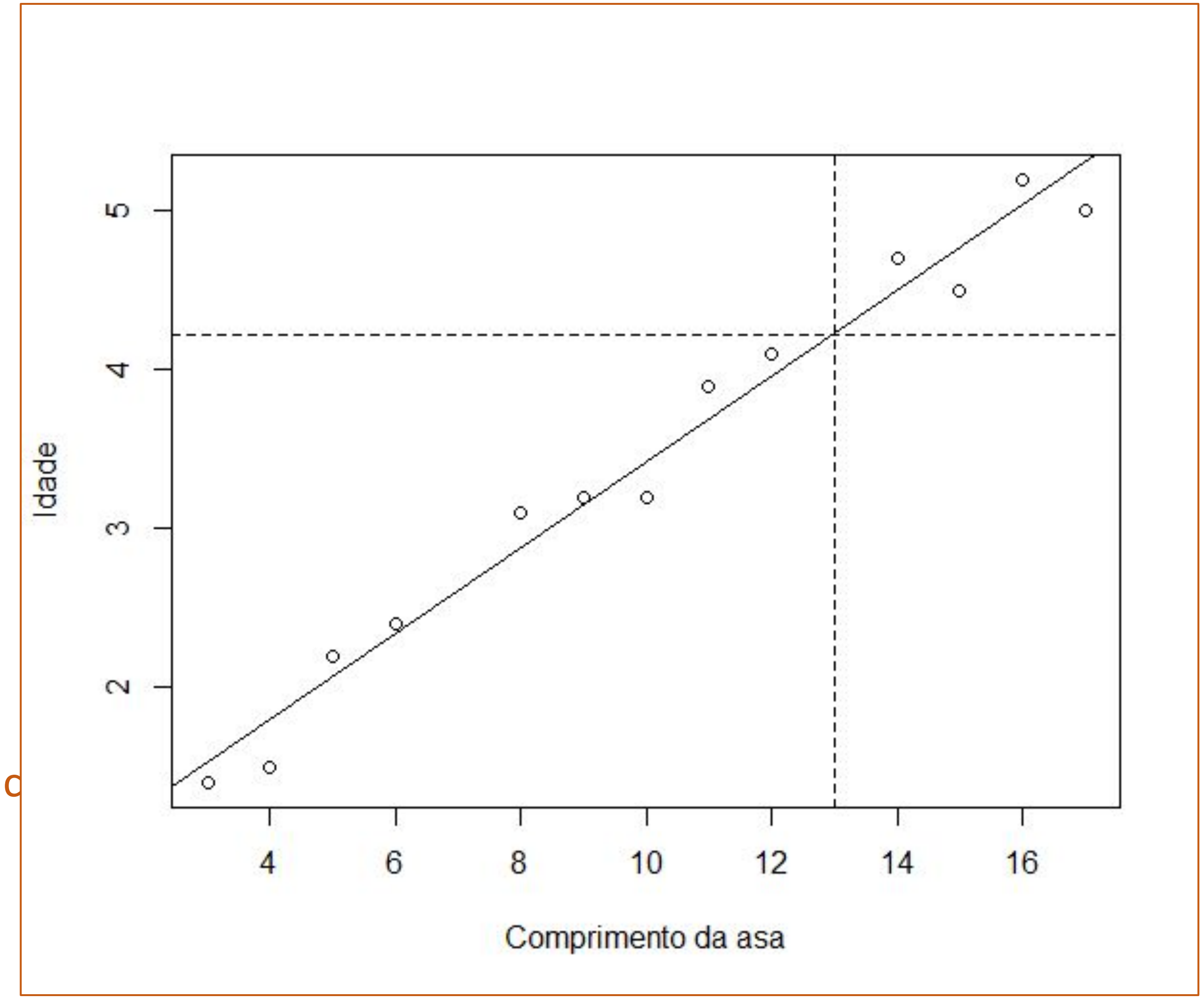
cm cm/dia dias

Regressão Linear Simples

- Calculando os valores preditos de Y

cm

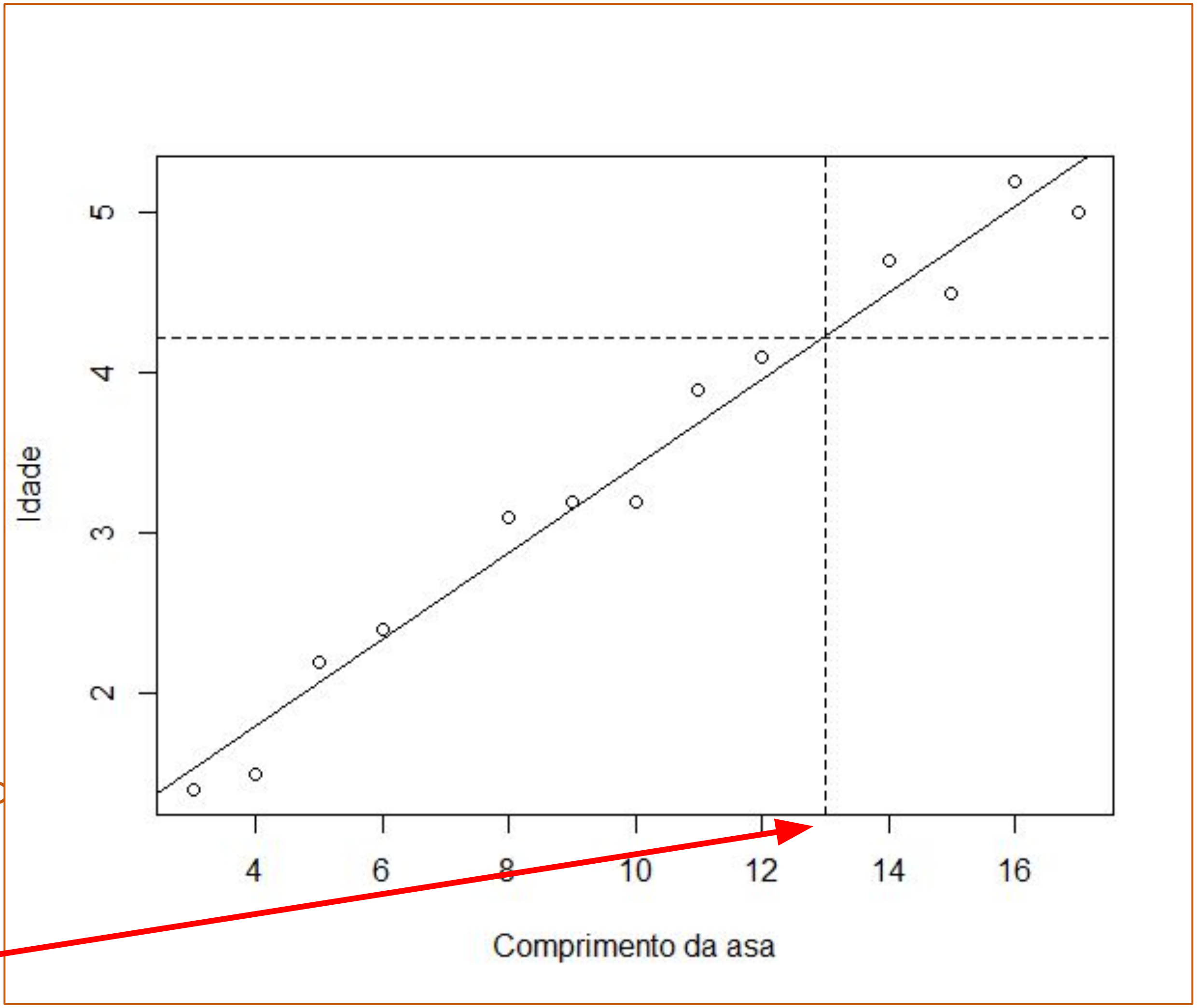
c



Regressão Linear Simples

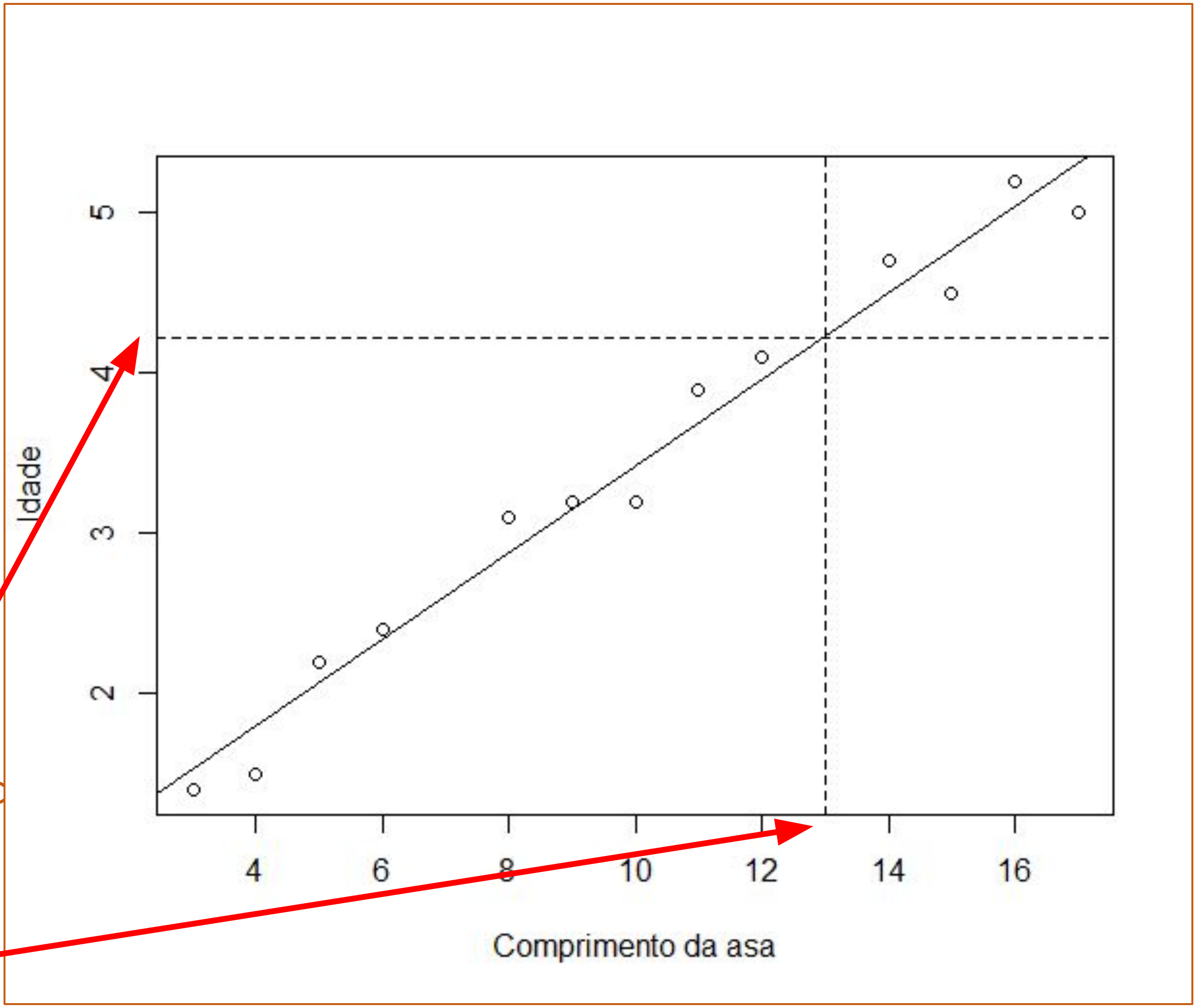
- Calculando os valores preditos de Y

cm



Regressão Linear Simples

- Calculando os valores preditos de Y



cm

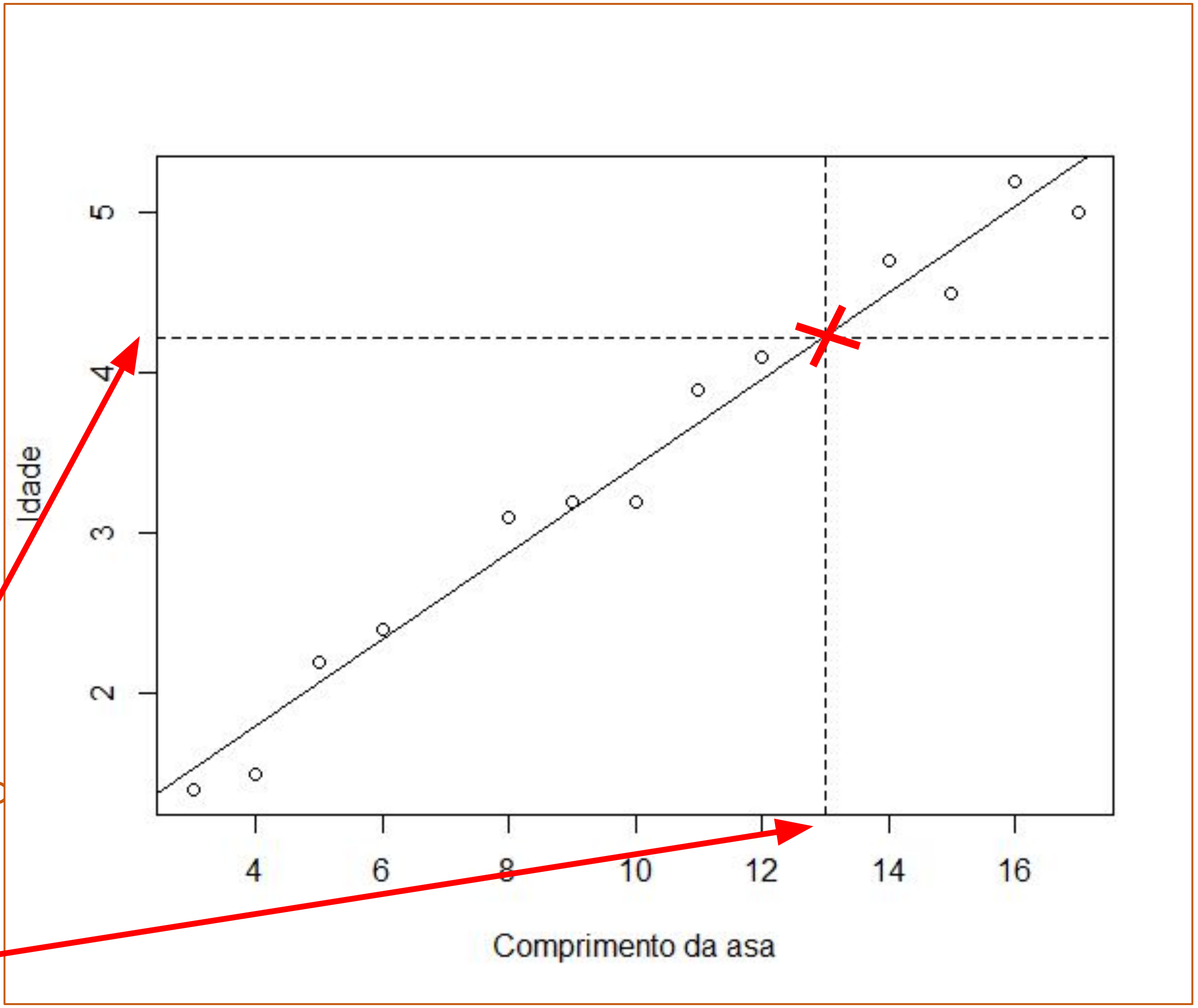
c

Regressão Linear Simples

- Calculando os valores preditos de Y

cm

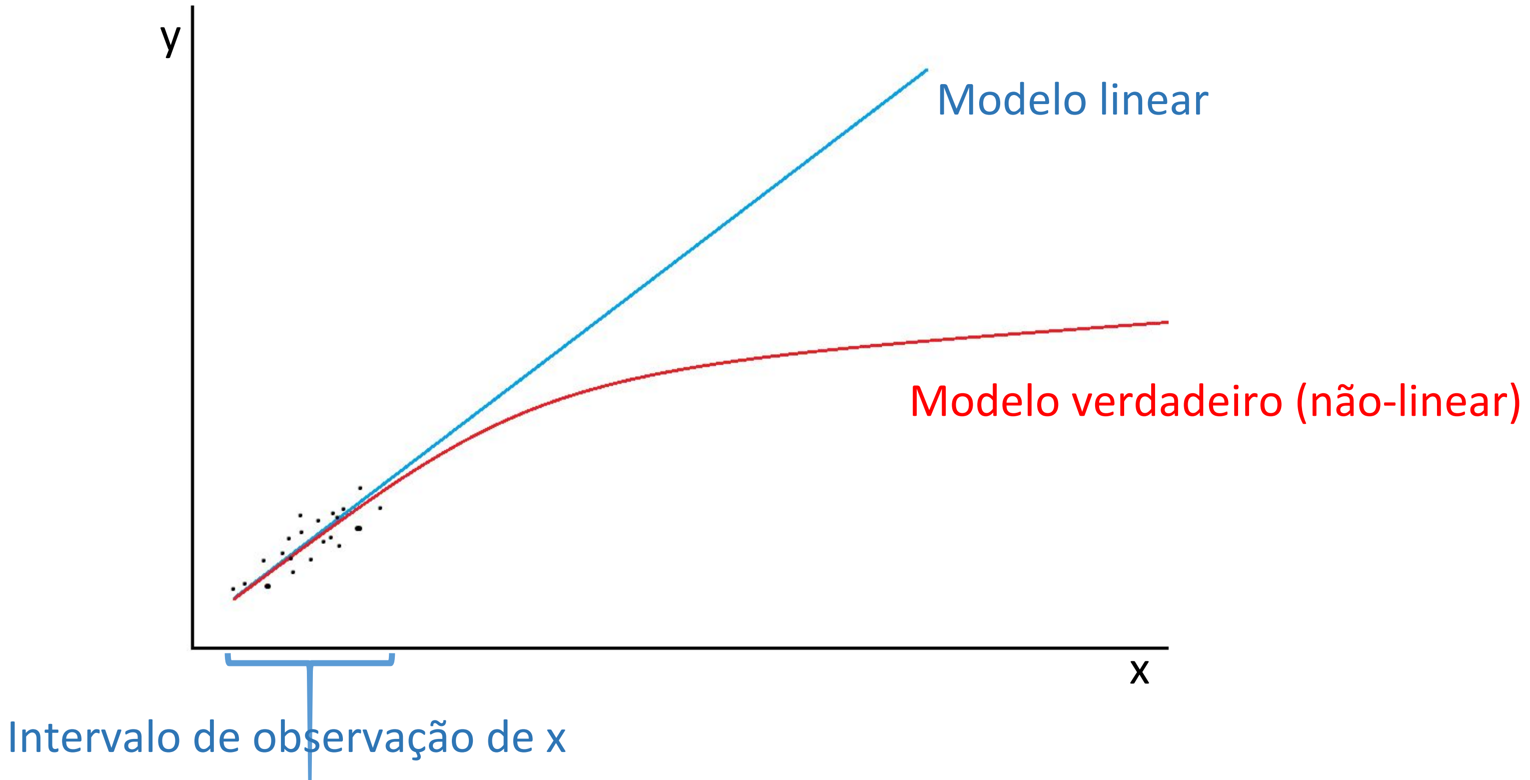
c



Regressão Linear Simples

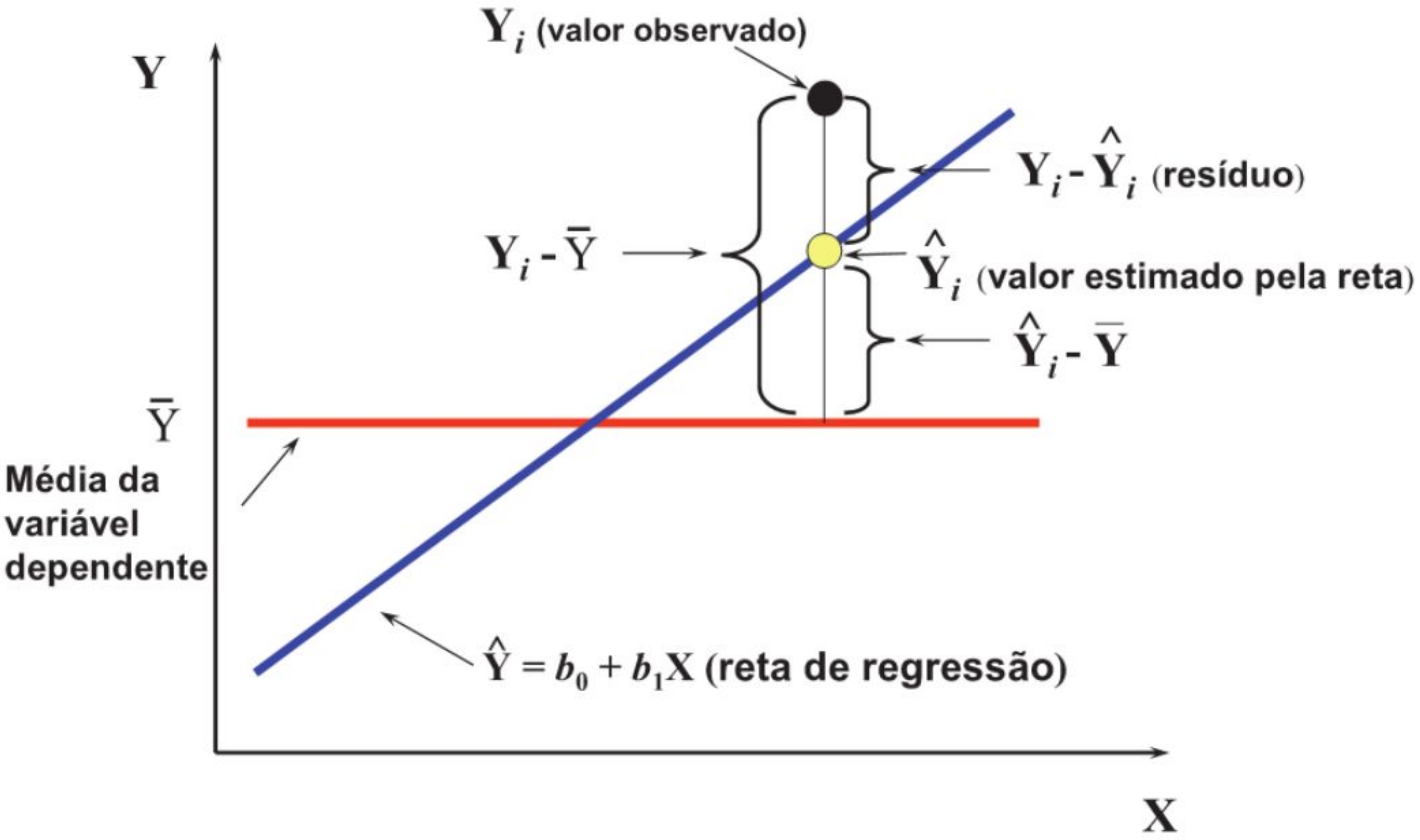
Cuidados que devemos ter para estimar os valores de Y

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

Decomposição do erro



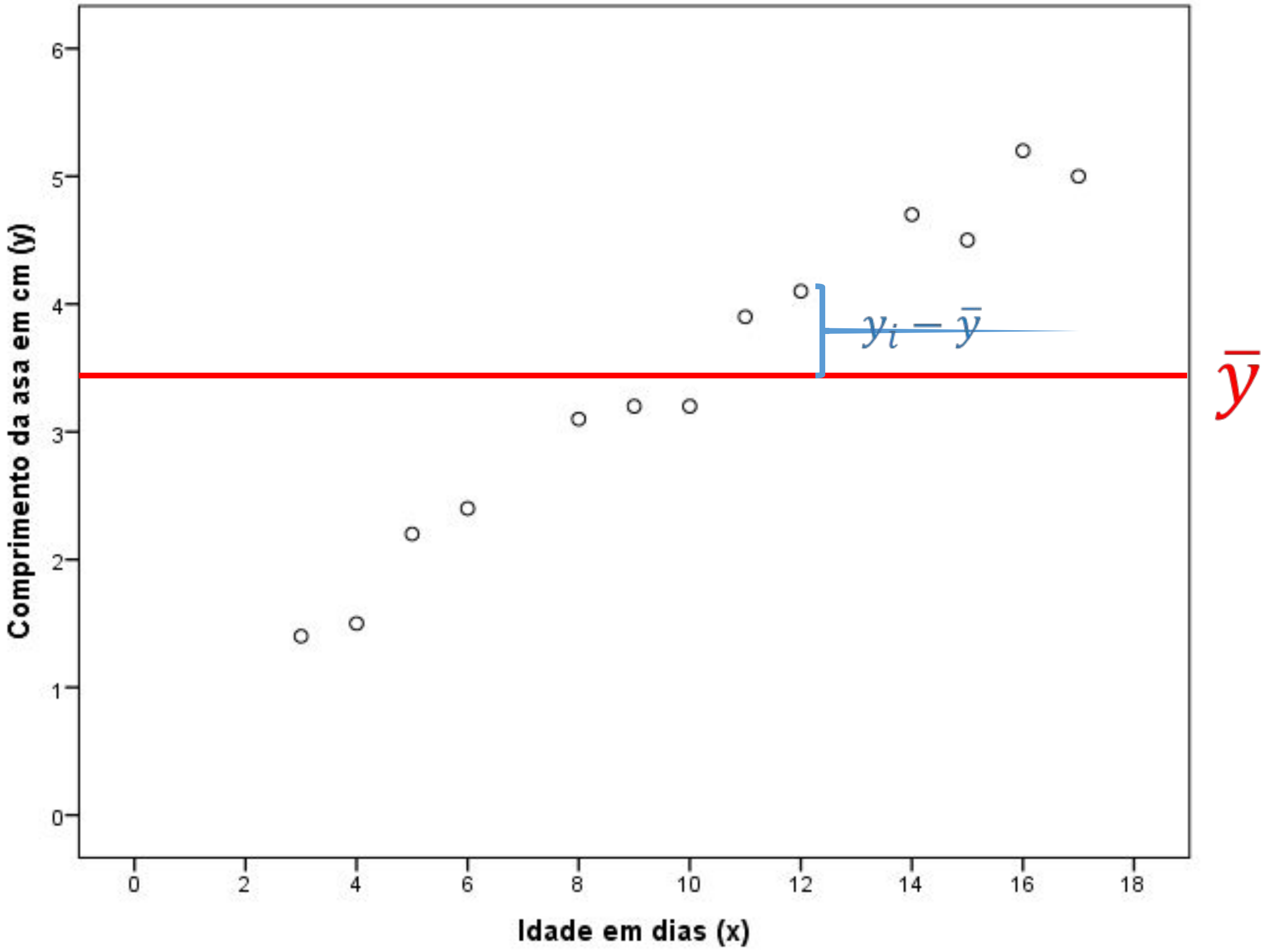
Regressão Linear Simples

Decomposição do erro

$$SQ_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

SQ_{total} -> é a soma dos quadrados dos desvios de y em relação a sua média ($\bar{y}$).

É uma medida da variação total da variável dependente (y).



SQ_{total} possui $n - 1$ gl.

Regressão Linear Simples

Decomposição do erro

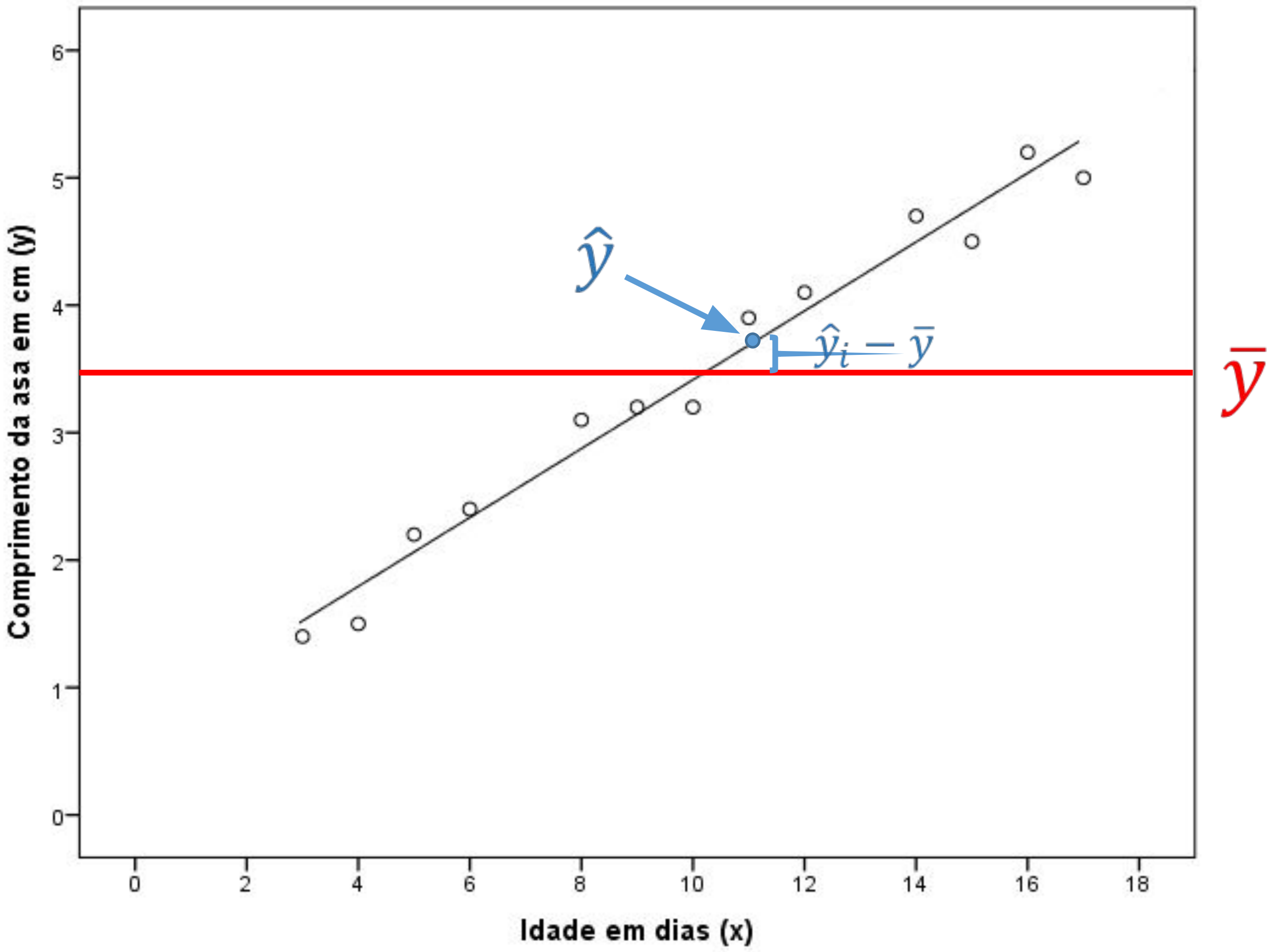
$$SQ_{regressão} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$SQ_{regressão}$ -> é a soma dos quadrados dos desvios entre a reta de regressão ($\hat{y}$) e a média ($\bar{y}$).

É uma medida da parcela da variação da variável dependente (y) que é explicada pela regressão, ou seja, pela variável (x).

$SQ_{regressão}$ possui k gl.

k -> é o número de variáveis independentes, no caso de regressão linear simples, é 1.



Regressão Linear Simples

Decomposição do erro

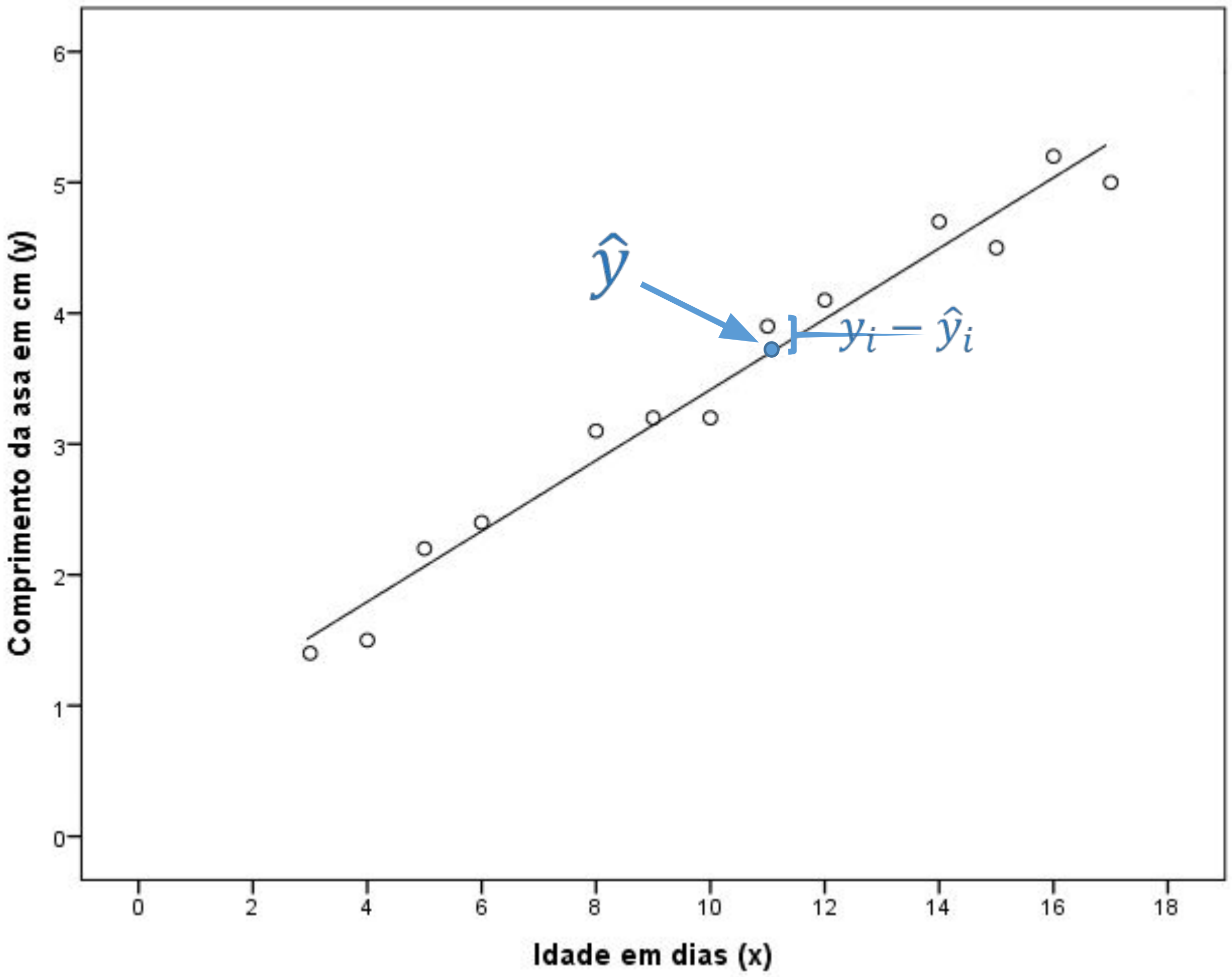
$$SQ_{residual} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$SQ_{residual}$ -> é a soma dos quadrados dos desvios de y em relação a reta de regressão ($\hat{y}$).

É uma medida da parcela da variação de y que não é explicada pela regressão, ou seja, pela variável (x).

$SQ_{residual}$ possui $n - k - 1$ gl.

k -> é o número de variáveis independentes, no caso de regressão linear simples, é 1.



Regressão Linear Simples

Decomposição do erro

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SQ_{total} \quad = \quad SQ_{regressão} \quad + \quad SQ_{residual}$$

Regressão Linear Simples

Análise de variância

Fonte	Soma dos quadrados (SQ)	Graus de Liberdade (g.l.)	Quadrados Médios (QM)	F	P_valor
Regressão	$SQ_{regressão}$	K	$MQ_{reg} = \frac{SQ_{reg}}{k}$	$F = \frac{MQ_{regressão}}{MQ_{residual}}$	P-valor
Erro	$SQ_{residual}$	n-k-1	$MQ_{residual} = \frac{SQ_{residual}}{n-k-1}$		
Total	SQ_{total}	n-1			

Regressão Linear Simples

Coeficiente de determinação

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

Teste t de Student

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

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

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

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(for          omprimento 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

Residu          d 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 $S_{\beta 0}$

β_1 $S_{\beta 1}$

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
```

β_0 $S_{\beta 0}$

β_1 $S_{\beta 1}$

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

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

Regressão Linear Simples

Coeficientes do modelo (Exemplo dos pardais)

```
> mod_exempl<-lm(`Comprimento da asa` ~ Idade, pardais)
> summary(mod_exempl)
Call:
lm(for          omprimento 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

Residu                d 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

$S_{\beta 0}$

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

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

β_1

$S_{\beta 1}$

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

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)

$$R_a^2 = 1 - \frac{\frac{SQresidual}{n-k-1}}{\frac{SQtotal}{n-1}}$$

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

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 ($X_1, X_2, \ldots, X_k$)
- 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

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \qquad \textit{i=1,n}$$

β_0 , β_1 , β_2 , ... , β_k são parâmetros do modelo

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

Regressão Linear Múltipla

Modelo estatístico

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

β_0 , β_1 , β_2 , ... , β_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 βi (i=1,...,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

$\beta 0$: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:

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

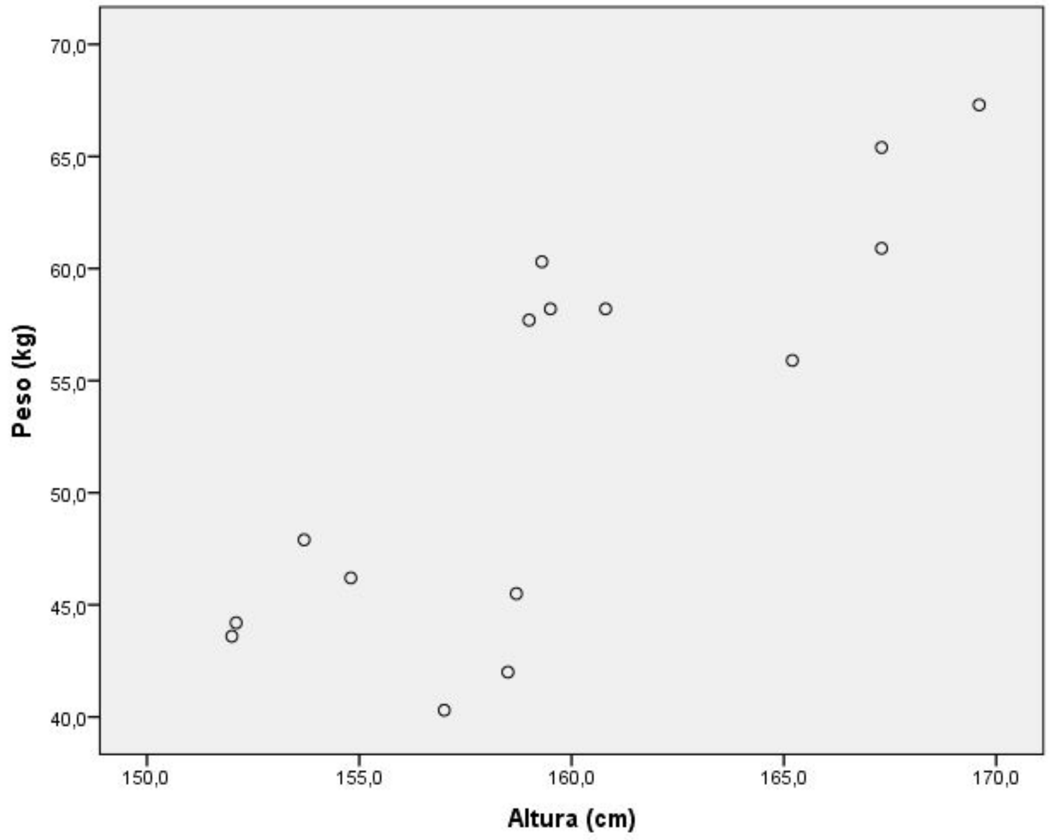
Regressão Linear Múltipla

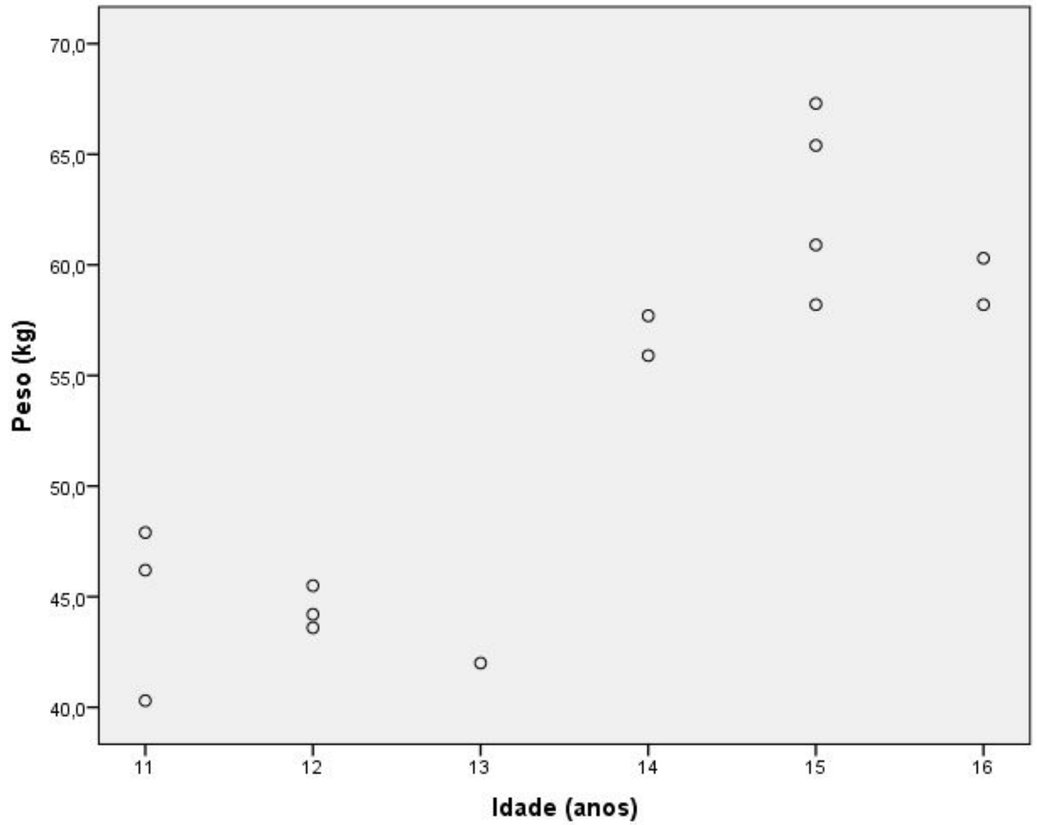
Exemplo: Dados de peso, idade e altura de 15 crianças:

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 **
```

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

```
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. 4ª. 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

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.

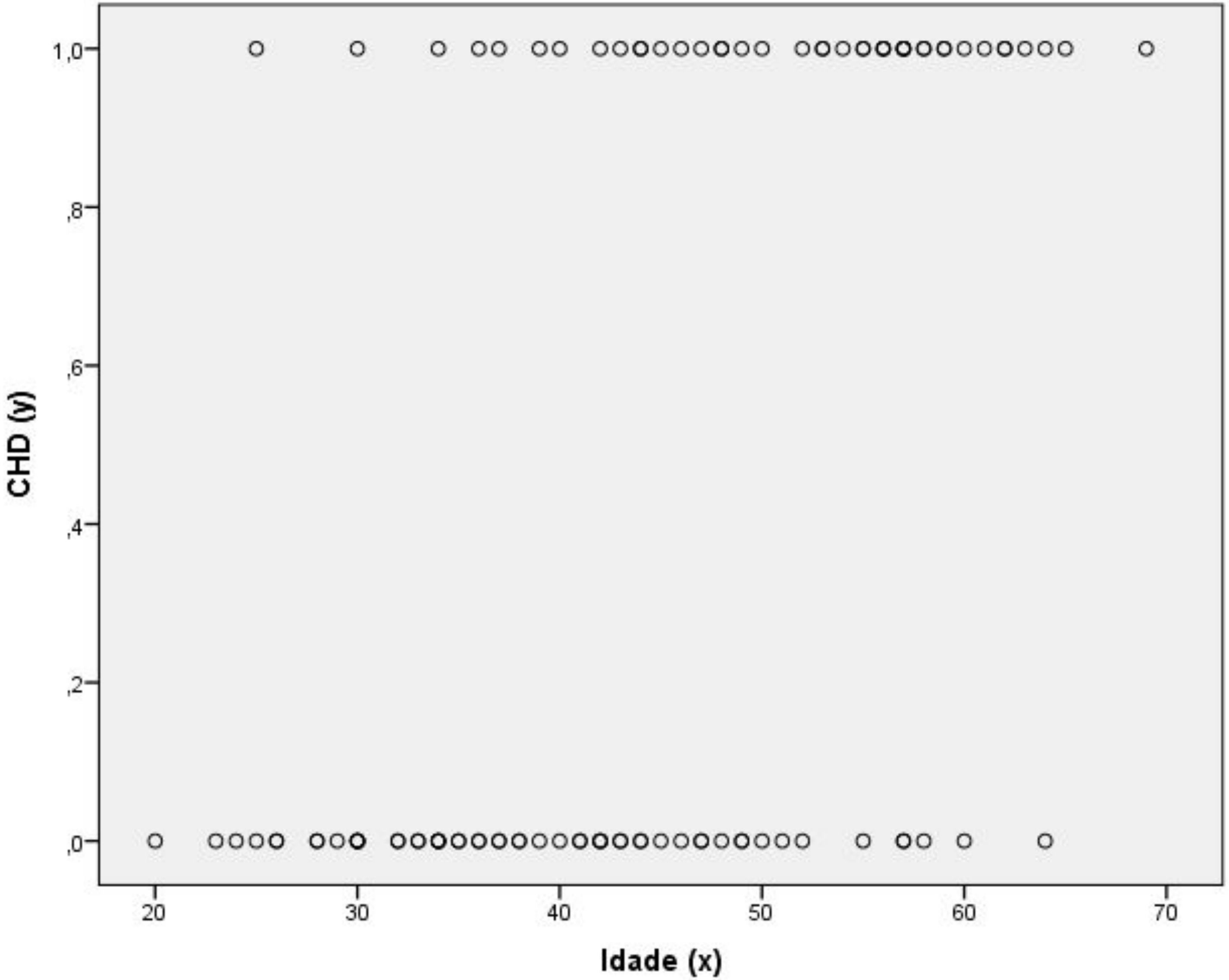
Regressão Logística

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

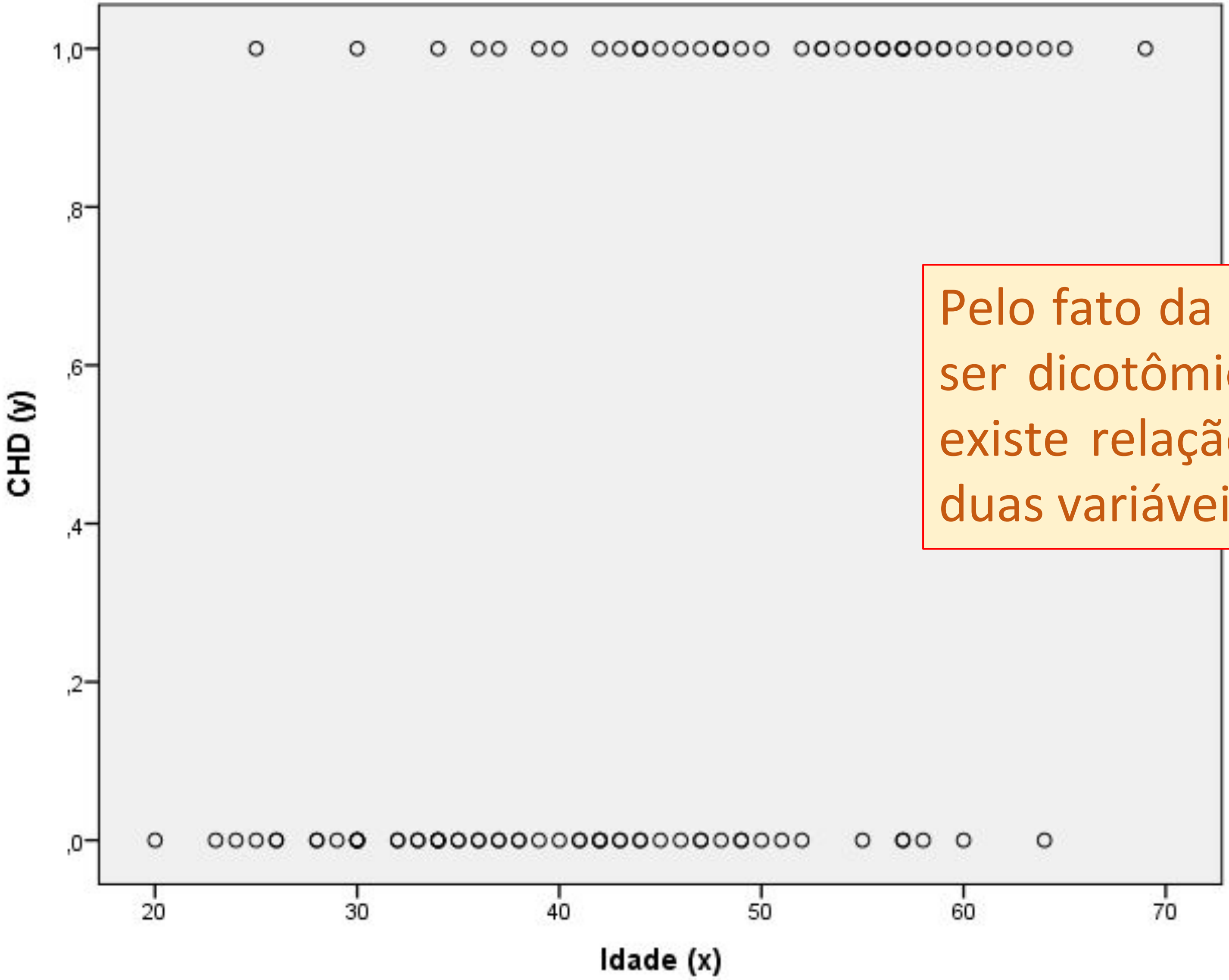
Exemplo: Usando os dados sobre presença ou ausência de doença coronária (CHD) e idade (AGE) de 100 indivíduos (Fonte: Hosmer & Lemeshow; 1989), desejamos analisar a relação entre a CHD e idade, tendo como variável dependente a CHD e como variável independente a idade.

Regressão Logística



Fonte: Hosmer & Lemeshow (1989)

Regressão Logística



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

Fonte: Hosmer & Lemeshow (1989)

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.

Fonte: Hosmer & Lemeshow (1989)

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%

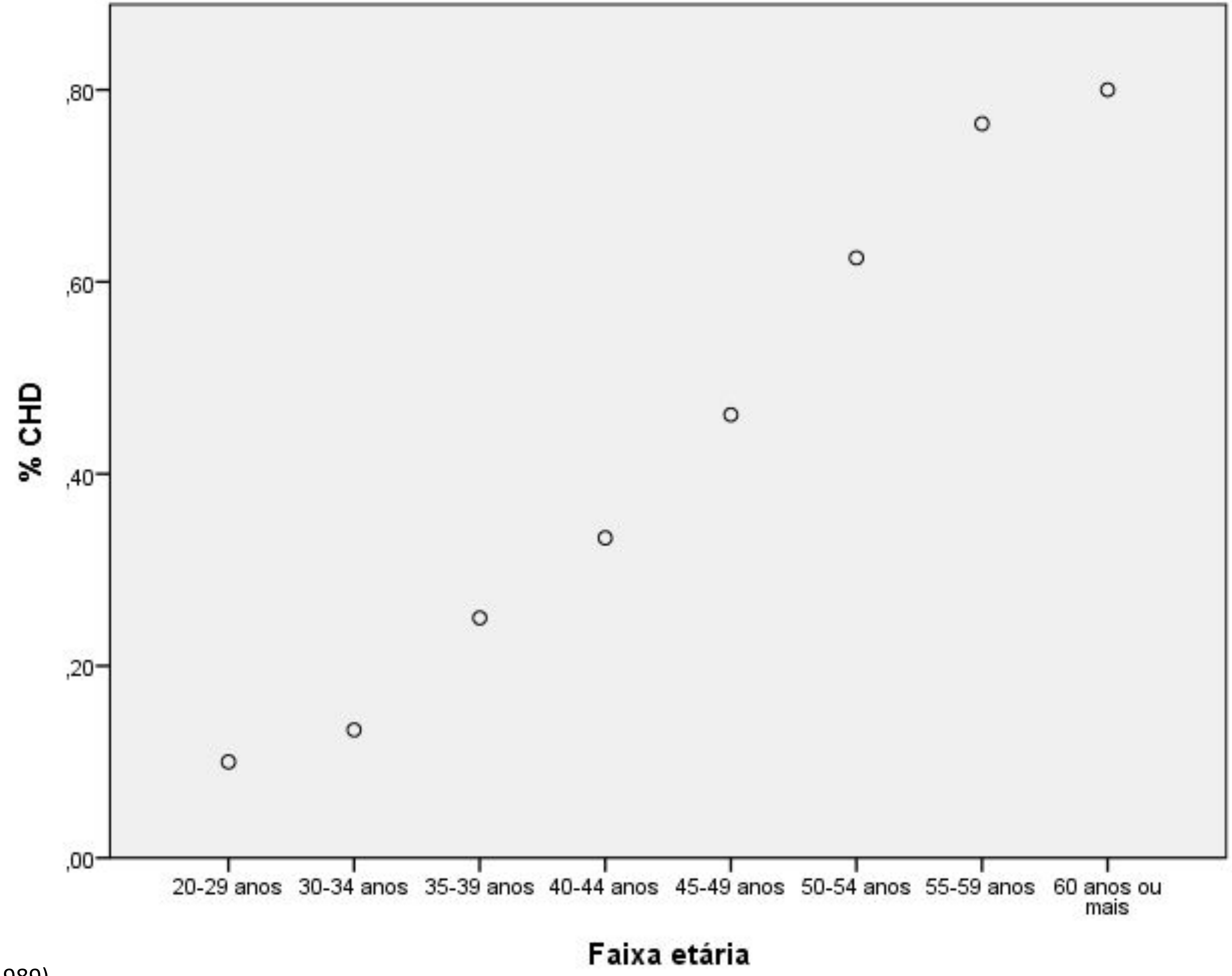
Fonte: Hosmer & Lemeshow (1989)

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%

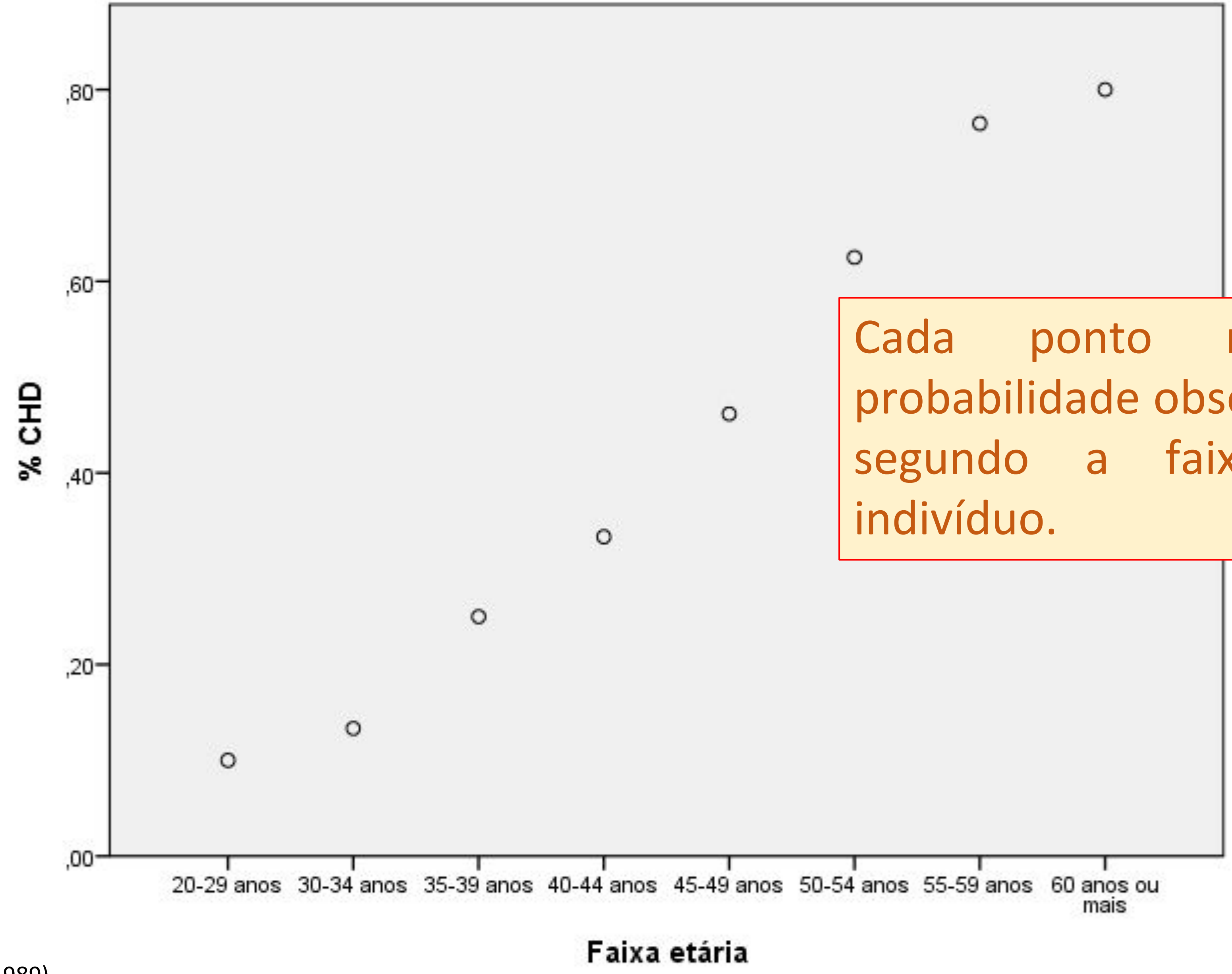
Fonte: Hosmer & Lemeshow (1989)

Regressão Logística



Fonte: Hosmer & Lemeshow (1989)

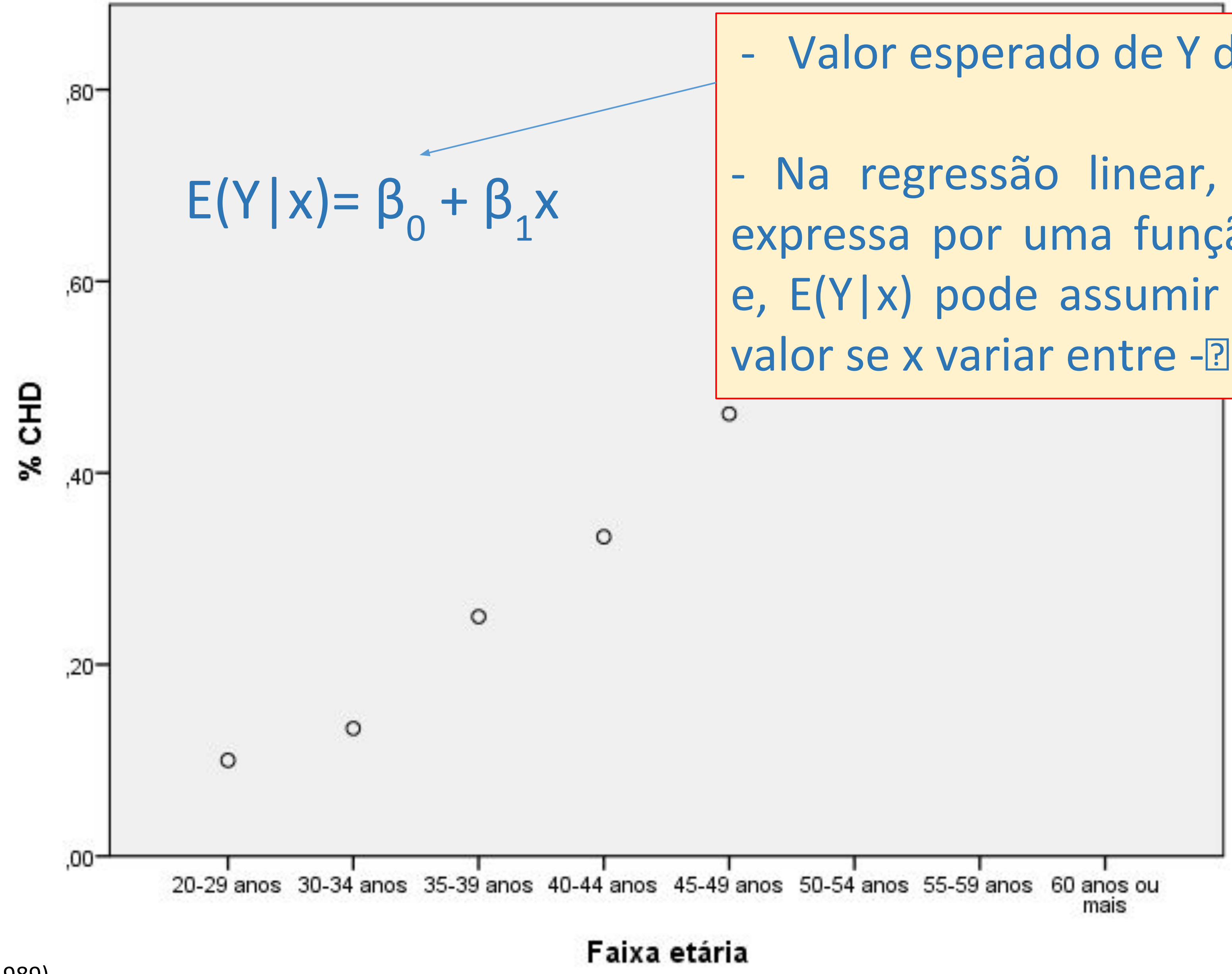
Regressão Logística



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

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística



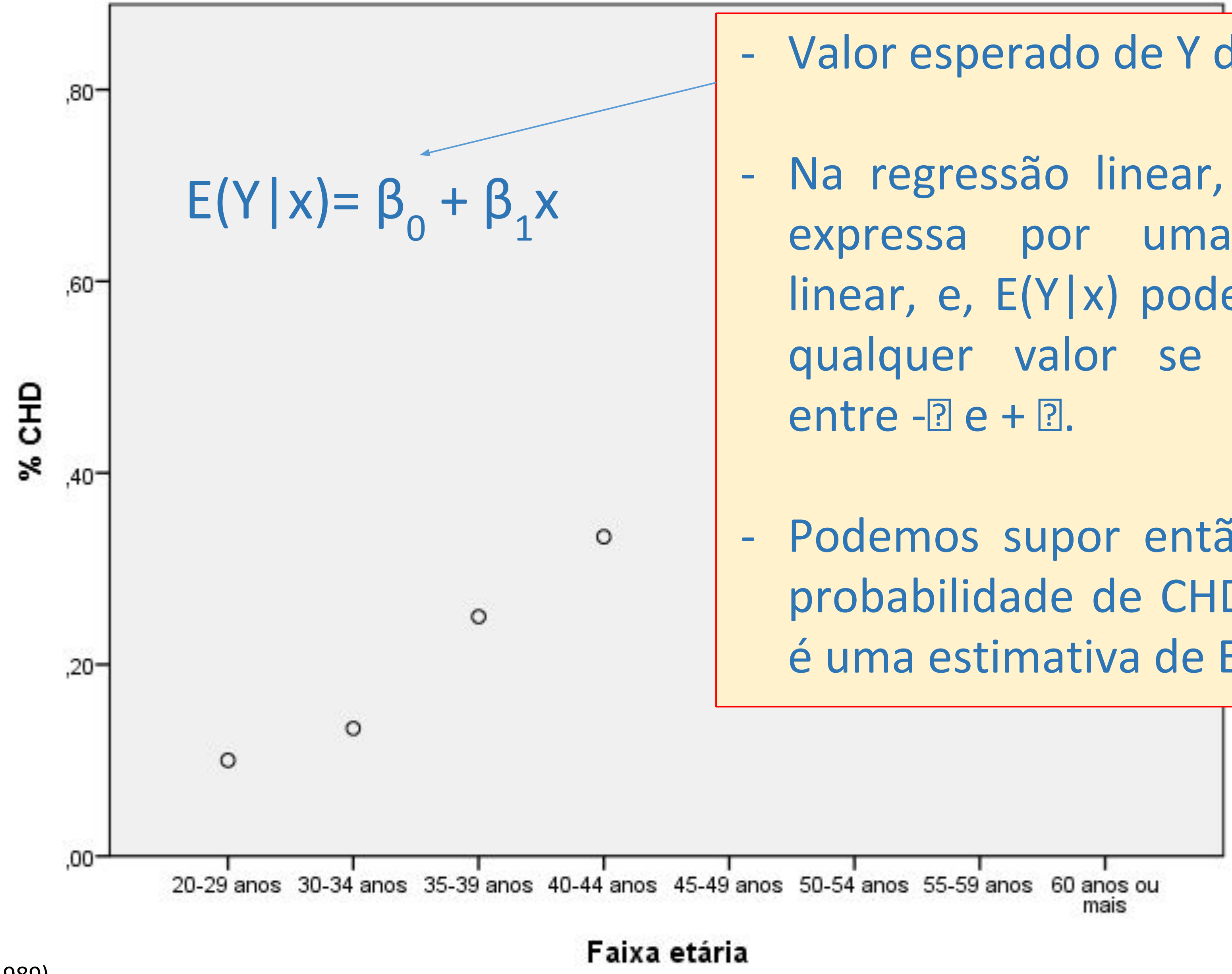
$E(Y|x) = \beta_0 + \beta_1 x$

- 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 $+\infty$.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

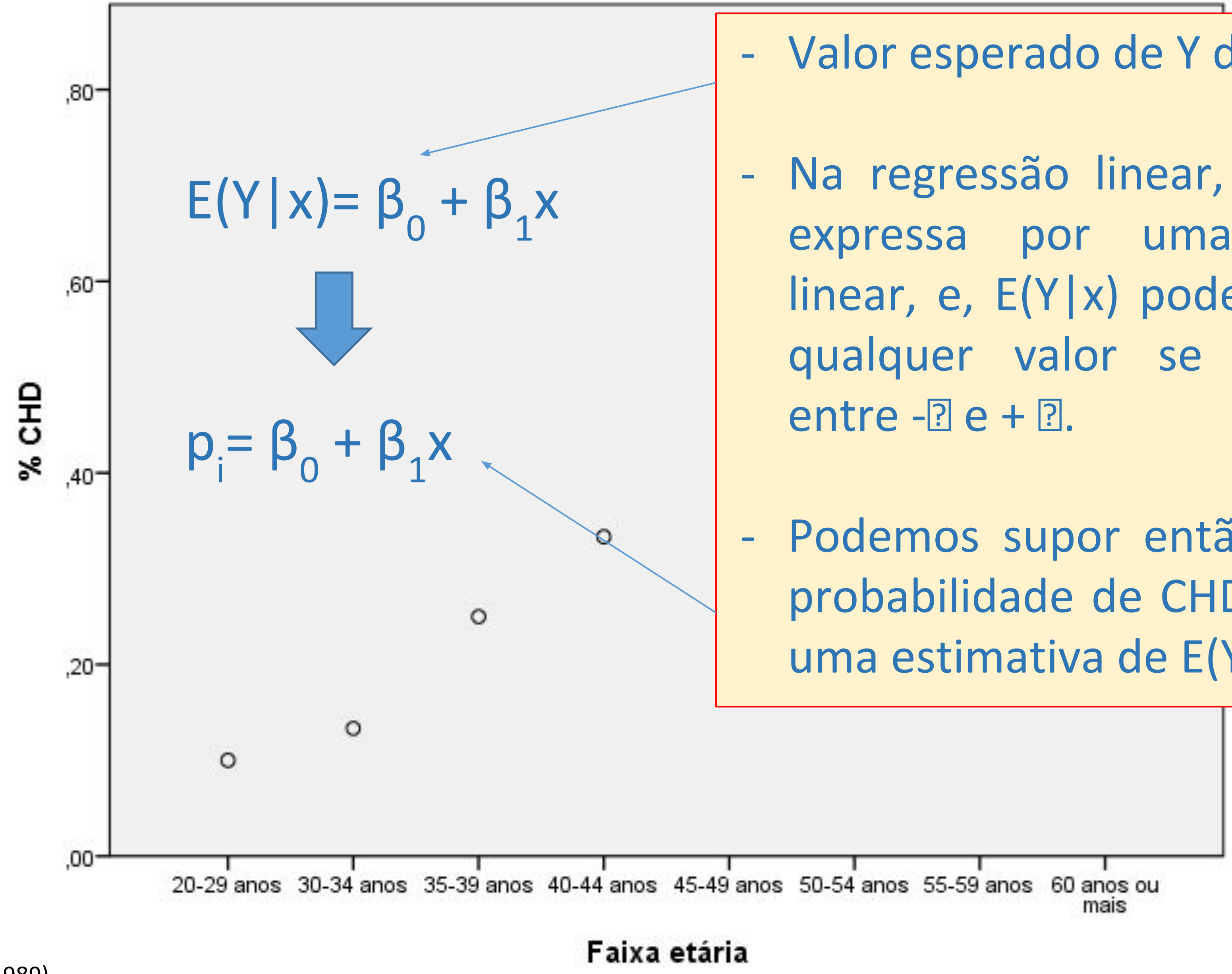


$E(Y|x) = \beta_0 + \beta_1 x$

- 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 $E(Y|x)$.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística



$$E(Y|x) = \beta_0 + \beta_1 x$$

$$p_i = \beta_0 + \beta_1 x$$

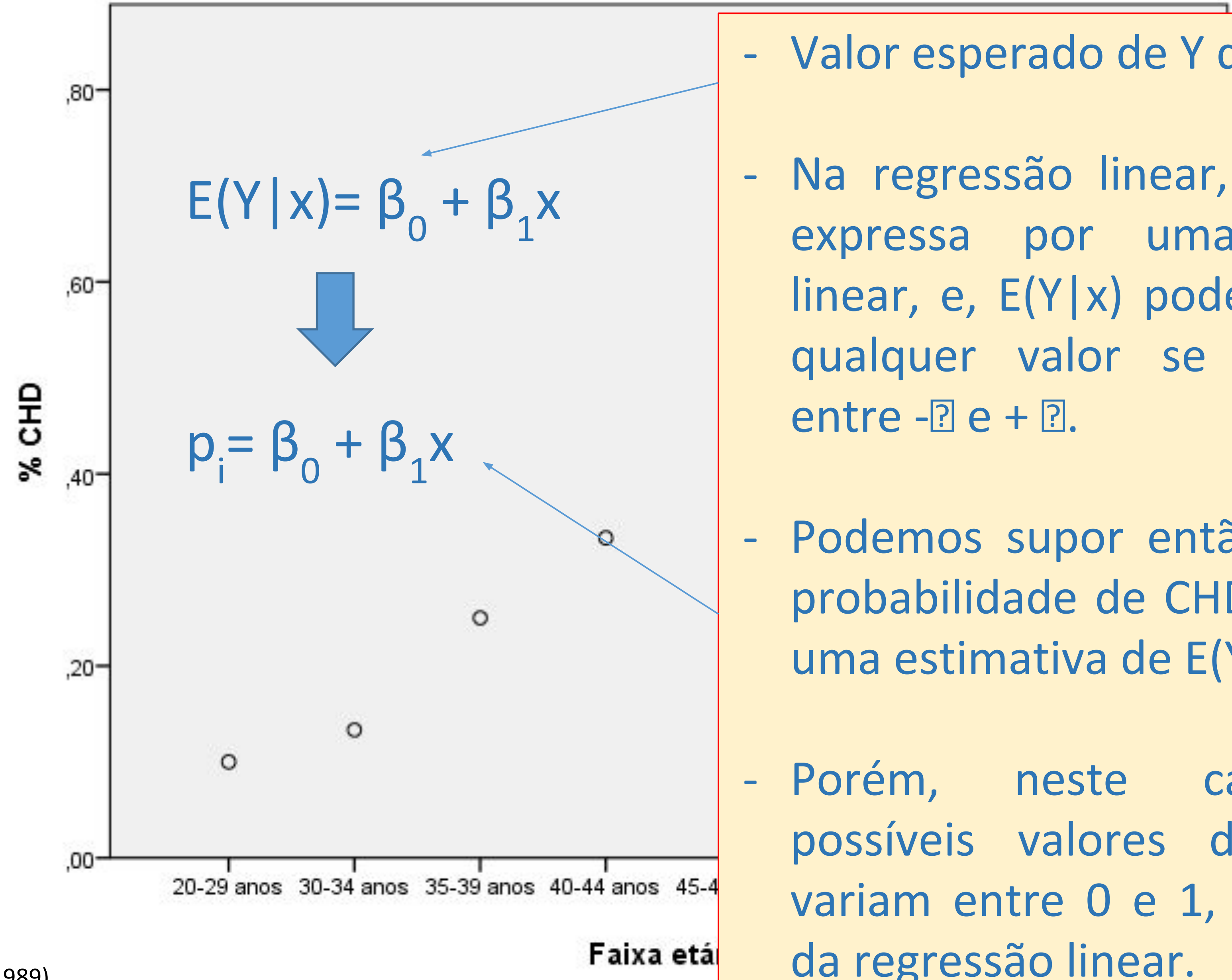
- Valor esperado de Y dado x.

- Na regressão linear, E(Y|x) é expressa por uma função linear, e, E(Y|x) pode assumir qualquer valor se x variar entre -∞ e + ∞.

- Podemos supor então, que a probabilidade de CHD (eixo Y) uma estimativa de E(Y|x).

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística



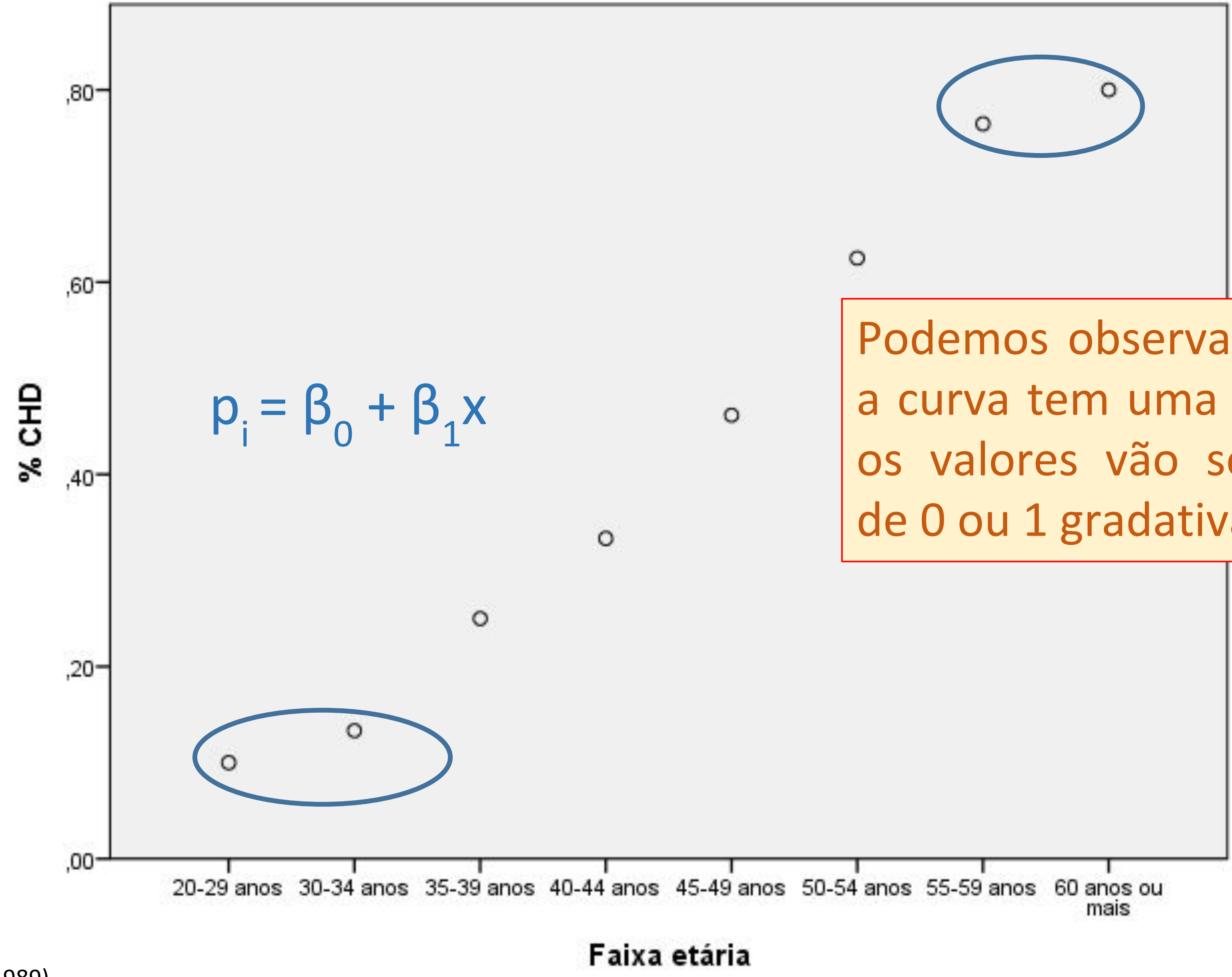
$$E(Y|x) = \beta_0 + \beta_1 x$$

$$p_i = \beta_0 + \beta_1 x$$

- 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.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística



$p_i = \beta_0 + \beta_1 x$

Podemos observar também, que a curva tem uma forma de "S" e os valores vão se aproximando de 0 ou 1 gradativamente.

Fonte: Hosmer & Lemeshow (1989)

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.

Fonte: Hosmer & Lemeshow (1989)

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.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Doença Coronária (DC)	Faixa etária >=55 (1)	 <55 (0)	Total
Sim	21	22	43
Não	6	51	57
Total	27	73	100

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Doença Coronária (DC)	Faixa etária >=55 (1)	<55 (0)	Total
Sim	21	22	43
Não	6	51	57
Total	27	73	100

Probabilidade de DC na faixa etária >=55 (1):

$$p_1 = \frac{21}{27} = 0{,}78$$

Probabilidade de DC na faixa etária <55 (0):

$$p_0 = \frac{22}{73} = 0{,}30$$

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Doença Coronária (DC)	Faixa etária >=55 (1)	<55 (0)	Total
Sim	21	22	43
Não	6	51	57
Total	27	73	100

Probabilidade de DC na faixa etária >=55 (1):

$$p_1 = \frac{21}{27} = 0,78$$

Probabilidade de DC na faixa etária <55 (0):

$$p_0 = \frac{22}{73} = 0,30$$

Chance de DC na faixa etária >=55 (1):

$$odds_1 = \frac{\frac{21}{27}}{\frac{6}{27}} = \frac{21}{6} = 3,5$$

Chance de DC na faixa etária <55 (0):

$$odds_0 = \frac{\frac{22}{73}}{\frac{51}{73}} = \frac{22}{51} = 0,43$$

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Relação entre probabilidade e odds:

$$odds = \frac{p}{1-p} \quad \text{ou} \quad p = \frac{odds}{1+odds}$$

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Probabilidade	Chance (odds)
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Probabilidade	Chance (odds)
0.1	0.11
0.2	0.25
0.3	0.43
0.4	0.67
0.5	1.00
0.6	1.50
0.7	2.33
0.8	4.00
0.9	9.00

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

A vantagem da odds neste caso é que não esta limitada entre 0 e 1, como a probabilidade.

Se a probabilidade da doença é 0,20 em um grupo A e 0,40 no grupo B, eu poderia dizer que a probabilidade da doença em A é o dobro da probabilidade em B.

Porém, se a probabilidade da doença no grupo A fosse 0,70, é impossível observar uma probabilidade que seja o dobro deste valor.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Isso não ocorre com a odds, pois a odds correspondente a um probabilidade de 0,70 é 2,33.

O dobro desta odds seria de 4,66, que corresponde a uma probabilidade de 0,82.

A razão entre a duas odds se chama Razão de Chance (odds ratio -> OR), neste caso a OR=2.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Voltando ao exemplo da doença coronária:

Doença Coronária (DC)	Faixa etária >=55 (1)	<55 (0)	Total
Sim	21	22	43
Não	6	51	57
Total	27	73	100

$$OR = \frac{\frac{21}{6}}{\frac{22}{51}} = \frac{3,50}{0,43} = 8,14$$

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Voltando ao exemplo da doença coronária:

Doença Coronária (DC)	Faixa etária >=55 (1)	<55 (0)	Total
Sim	21	22	43
Não	6	51	57
Total	27	73	100

$$OR = \frac{\frac{21}{6}}{\frac{22}{51}} = \frac{3,50}{0,43} = 8,14$$

A chance de ocorrência de doença coronária entre indivíduos com 55 anos ou mais foi 8 vezes mais que a chance de ocorrência de DC entre os mais jovens.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Transformando a probabilidade da seguinte forma, resolvemos o problema do limite entre 0 e 1:

- Ao utilizarmos a odds, resolvemos o problema do limite superior;
- Ao usarmos o logaritmo da odds, resolvemos o problema do limite inferior

O modelo logístico é a relação linear desta transfomação com os preditores.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Sendo assim,

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

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

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Sendo assim,

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

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

Denominado: logito ou log-odds

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Sendo assim,

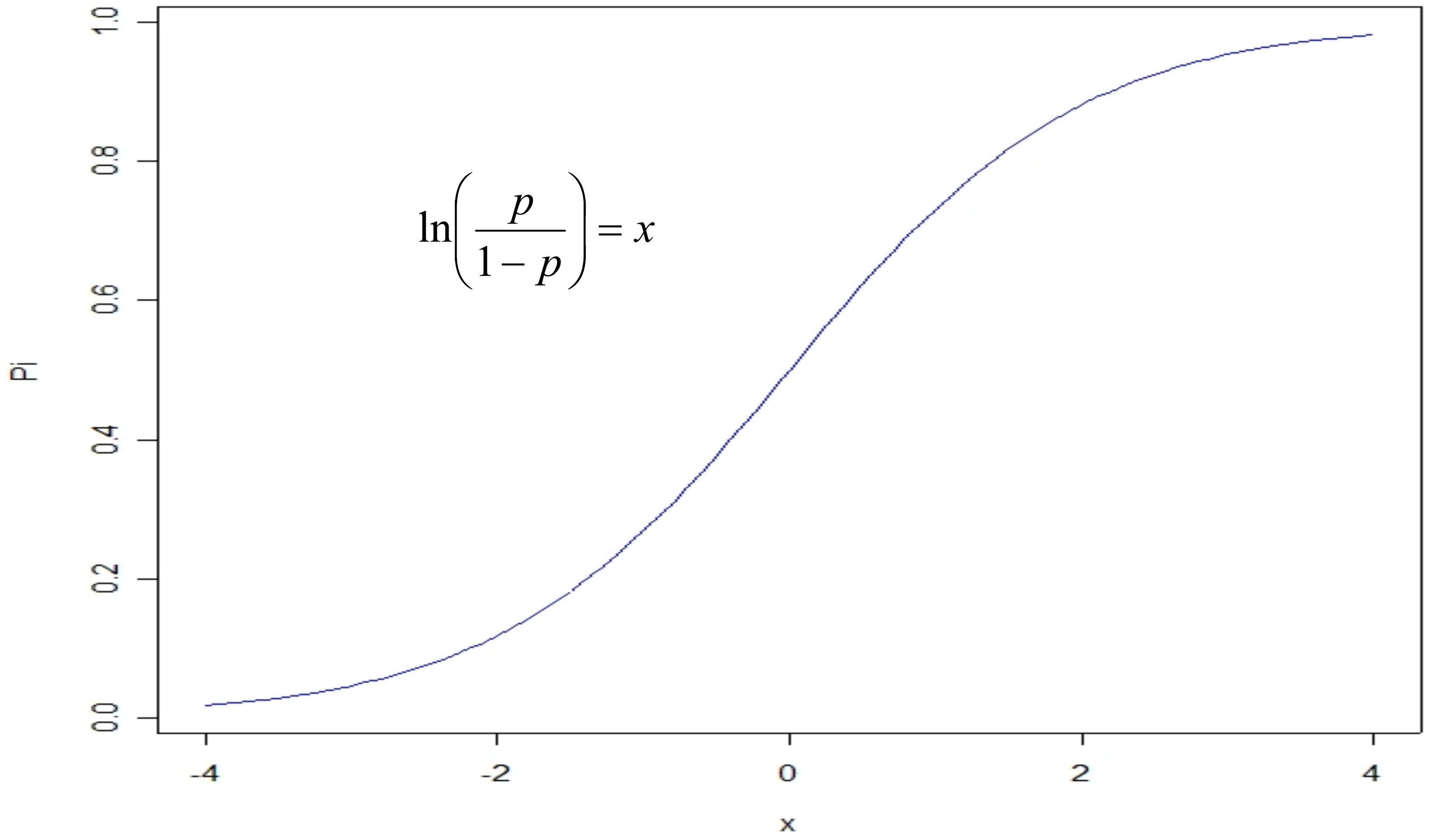
$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$$

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

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

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística



Fonte: Hosmer & Lemeshow (1989)

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

Fonte: Hosmer & Lemeshow (1989)

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

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

Fonte: Hosmer & Lemeshow (1989)

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

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

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Razão de chance (OR):

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

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Razão de chance (OR):

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

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

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Estimação dos parâmetros

O método dos mínimos quadrados, utilizado nos modelos lineares, não gera estimativas eficientes dos parâmetros quando a variável dependente é dicotômica.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Estimação dos parâmetros

O método utilizado para estimar os parâmetros no modelo de regressão logística é o da Máxima Verossimilhança.

O método da Máxima Verossimilhança fornece valores para os parâmetros que maximizam a probabiliadae de se obter um determinado conjunto de dados.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Este método equivale ao método dos mínimos quadrados sob o modelo de regressão linear (quando os resíduos são normalmente distribuídos).

Consiste em maximizar a função de verossimilhança. Essa função expressa a probabilidade dos dados observados em função dos parâmetros desconhecidos.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Os estimadores de máxima verossimilhança são escolhidos para serem os que maximizam esta função.

Sendo assim, os estimadores resultantes são aqueles que concordam mais estreitamente com o conjunto de dados observados.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Se os valores de Y são codificados como 0 ou 1, então, p_i da equação abaixo fornece os valores (para qualquer valor arbitrário de β_0 e β_1) a probabilidade de Y=1 dado x, ou seja, $E[Pr(Y=1|x)] = p_i$ e $E[Pr(Y=0|x)] = 1-p_i$.

$$p_i = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$$

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Assim, para os pares (x_i, y_i), onde $y_i = 1$ a contribuição para a função de verossimilhança é p_i e, para os pares onde $y_i = 0$ a contribuição para a função de verossimilhança é $1 - p_i$. Onde p_i denota o valor de p computado para x_i .

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Uma maneira conveniente de expressar a contribuição a contribuição, para a função de verossimilhança, do par (x_i, y_i) é através da equação:

$$W(x_i) = p_i^{y_i} * [1 - p_i]^{1-y_i}$$

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Uma maneira conveniente de expressar a contribuição a contribuição, para a função de verossimilhança, do par (x_i, y_i) é através da equação:

$$W(x_i) = p_i^{y_i} * [1 - p_i]^{1-y_i}$$

Quando $y_i = 1$ a contribuição para a função é p_i.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Uma maneira conveniente de expressar a contribuição a contribuição, para a função de verossimilhança, do par (x_i, y_i) é através da equação:

$$W(x_i) = p_i^{y_i} * [1 - p_i]^{1-y_i}$$

Quando $y_i = 0$ a contribuição para a função é $1\text{-}p_i$.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Uma vez que assumimos as observações como independentes, a função de verossimilhança é obtida pelo produto dos termos de $W(x_i)$:

$$l(\beta) = \prod_{i=1}^{n} W(x_i)$$

Pelo método da máxima verossimilhança, a estimativa de β é o valor que maximiza esta função.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Entretanto, é mais fácil, matematicamente, trabalhar com o Log da função de verossimilhança:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \, ln[p_i] + (1 - y_i) \, ln[1 - p_i]\}$$

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Entretanto, é mais fácil, maten ... m o Log da função de verossimilhança:

$$p_i = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$$

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \, ln[p_i] + (1 - y_i) \, ln[1 - p_i]\}$$

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Entretanto, é mais fácil, matematicamente, trabalhar com o Log da função de verossimilhança:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \, ln[p_i] + (1 - y_i) \, ln[1 - p_i]\}$$

Para achar o valor de β que maximiza L(β), calcula-se a derivada em relação a β e iguala-se a zero.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Estas equações são as seguintes:

$$\sum_{i=1}^{n} [y_i - p_i] = 0$$

$$\sum_{i=1}^{n} x_i [y_i - p_i] = 0$$

E são chamadas de equações de verossimilhança.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

Devido ao componente não linear destas equações, dificilmente a solução é obtida a não ser a partir de métodos iterativos.

Um dos métodos mais utilizados é o de Newton –Raphson.

Seja U(β), denominado de score, vetor com as derivadas de primeira ordem e I(β), denominado de Hessian, o vetor com as derivadas de segunda ordem.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Método de Máxima Verossimilhança

O algoritmo de Newton Raphson é:

$$\beta_{j+1} = \beta_j - I^{-1}(\beta_j)U(\beta_j)$$

Onde I^{-1} é a inversa de I

$$\beta_{j+1} - \beta_j < 0{,}0001$$

Fonte: Hosmer & Lemeshow (1989)

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);

Fonte: Hosmer & Lemeshow (1989)

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
```

Fonte: Hosmer & Lemeshow (1989)

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
```

Fonte: Hosmer & Lemeshow (1989)

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

β_1

Fonte: Hosmer & Lemeshow (1989)

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

$SE(\beta0)$

$SE(\beta1)$

β_1

Fonte: Hosmer & Lemeshow (1989)

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
```

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

$OR = Exp(\beta_0)$

```
> 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
```

Fonte: Hosmer & Lemeshow (1989)

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 = $Exp(\beta_0)$

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.

Fonte: Hosmer & Lemeshow (1989)

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 = Exp(β_0)

IC(95%) da OR

Exp($\beta_1 \pm z_{1-\alpha/2}$ * SE(β_1))

Exp(2,094 ±(1,96 * 0,529))

Fonte: Hosmer & Lemeshow (1989)

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 = Exp(β_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).

Fonte: Hosmer & Lemeshow (1989)

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 = Exp(β_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.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Teste de Wald

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

W tem distribuição Normal Padrão –> 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.

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

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

Call: glm(formula = CHD ~ fxet_55, family = bi

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

$H_0: \beta_i = 0$

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

Fonte: Hosmer & Lemeshow (1989)

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.

Fonte: Hosmer & Lemeshow (1989)

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).

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Razão de Verossimilhança

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

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

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Teste de Razão de Verossimilhança

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

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

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

Fonte: Hosmer & Lemeshow (1989)

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
```

Fonte: Hosmer & Lemeshow (1989)

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

Fonte: Hosmer & Lemeshow (1989)

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 Fis
```

D *do modelo só com a constante*

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

Fonte: Hosmer & Lemeshow (1989)

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
```

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

$\boxed{D \textit{ do modelo só com a constante}}$

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

$\boxed{D \textit{ do modelo após a inclusão da variável faixa etária } (x)}$

Fonte: Hosmer & Lemeshow (1989)

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 ***
```

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

$\boxed{D\ do\ modelo\ só\ com\ a\ constante}$

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

```
(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
```

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

$\boxed{D\ do\ modelo\ após\ a\ inclusão\ da\ variável\ faixa\ etária\ (x)}$

Fonte: Hosmer & Lemeshow (1989)

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 ***

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

(Dispersion parameter for

$D = 18{,}704$ $\chi 21gl = 3{,}84$ $p\text{-valor} < 0{,}001$

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

Fonte: Hosmer & Lemeshow (1989)

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 [$-2ln(verossimilhança\ do\ modelo\ com\ x$]

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

Fonte: Hosmer & Lemeshow (1989)

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
```

Fonte: Hosmer & Lemeshow (1989)

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
```

$p = 2, ou\ seja, dois\ parâmetros\ no\ modelo\ \beta_0\ e\ \beta_1$

D

$$\text{AIC} = 117{,}96 + 2\ (2) = 121{,}96$$

Fonte: Hosmer & Lemeshow (1989)

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 $Pr(Y_i=1)$:

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

Fonte: Hosmer & Lemeshow (1989)

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
```

Fonte: Hosmer & Lemeshow (1989)

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

β_1

Fonte: Hosmer & Lemeshow (1989)

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

β_0

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

β_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

```
> p1<-prob<-predict(mod_CHD, type = "response")
> table(p1)
p1
0.301369863013931       0.777777777777535
               73                      27
```

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

Fonte: Hosmer & Lemeshow (1989)

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

β_1

```
> p1<-prob<-predict(mod_CHD, type = "response")
> table(p1)
p1
0.301369863013931      0.777777777777535
               73                     27
```

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

Fonte: Hosmer & Lemeshow (1989)

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

β_0

Coefficients:

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

Probabilidade estimada de um indivíduo com 55 anos ou mais (x=1) ter doença coronária (y=1).

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

β_1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 136.66 on 99 degrees of freedom

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

Number of Fisher Scoring iterations: 4

Fonte: Hosmer & Lemeshow (1989)

Regressão Logística

Referências:

HOSMER, DW Jr. & LEMESHOW, S. Applied Logistic Regression.

1989. John Willey & Sons, Inc.