



# Correlação e regressão



Dados de mandíbulas de fetos  
Profa. Yana Borges

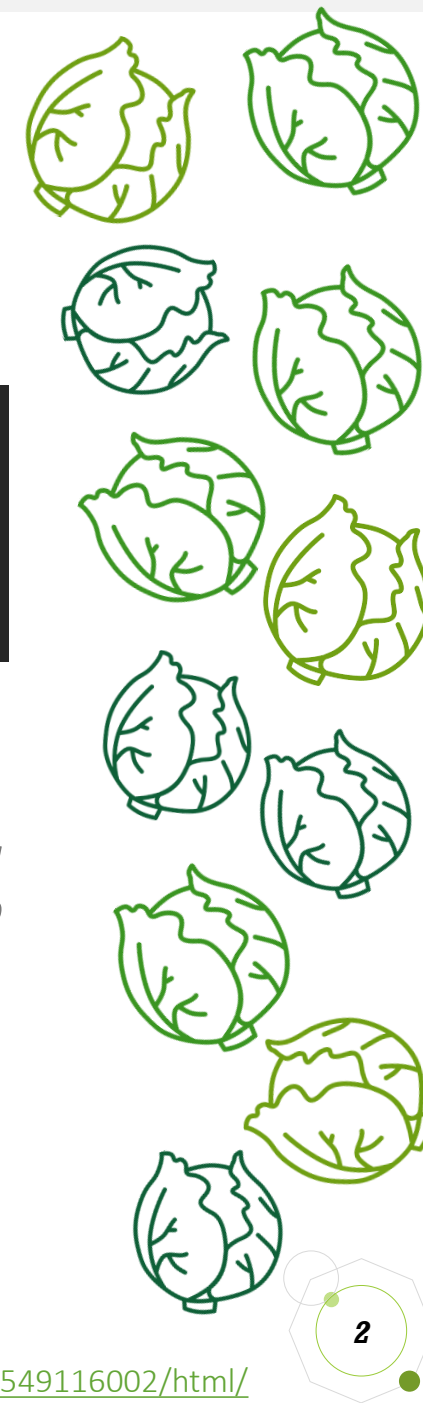


# Correlação

O objetivo do estudo correlacional é a determinação da força do relacionamento entre duas observações emparelhadas. Há muitos casos em que pode existir um relacionamento entre duas variáveis.

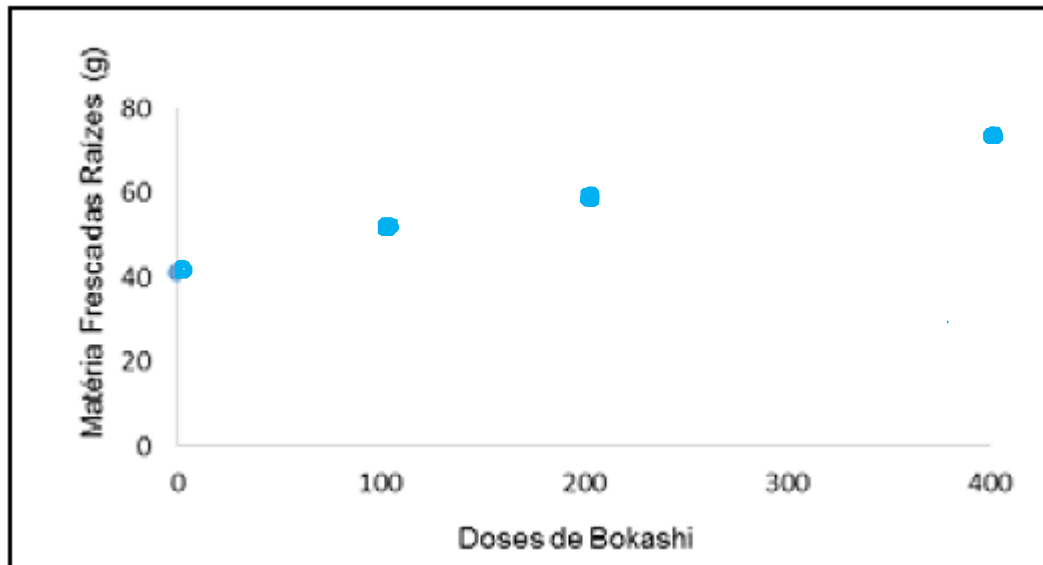
*Ao avaliar a incidência e a severidade da doença hérnia das crucíferas causada por Plasmodiophora brassicae W. em repolho (Brassica oleracea L. var. capitata) em solo submetido a tratamentos com doses de **biofertilizante tipo bokashi**, pergunta-se:*

*O teor de matéria fresca aumenta juntamente com a dosagem do bokashi<sup>1</sup>?*



# Correlação

Uma forma de visualizarmos se duas variáveis apresentam-se correlacionadas é através do *diagrama de dispersão*, onde os valores das variáveis são representados por pontos, num sistema cartesiano.



**Figura1:** Correlação entre os valores de massa de matéria fresca das raízes de repolho e doses de composto orgânico tipo bokashi<sup>1</sup>

A vantagem de se construir o diagrama de dispersão está em que, muitas vezes, sua simples observação já nos dá uma ideia bastante boa de como as duas variáveis se correlacionam.

1. <https://www.redalyc.org/journal/4675/467549116002/html/>

# Correlação linear

Os pontos obtidos, vistos em conjunto, formam uma elipse em diagonal. Podemos imaginar que, quanto mais fina for a elipse, mais ela se aproxima de uma reta. Dizemos, então, que a correlação de forma elíptica tem como “imagem” uma reta, sendo, por isso, denominada correlação linear.

## Correlação linear positiva

Quando os pontos do diagrama têm como imagem uma reta crescente.

## Correlação linear negativa

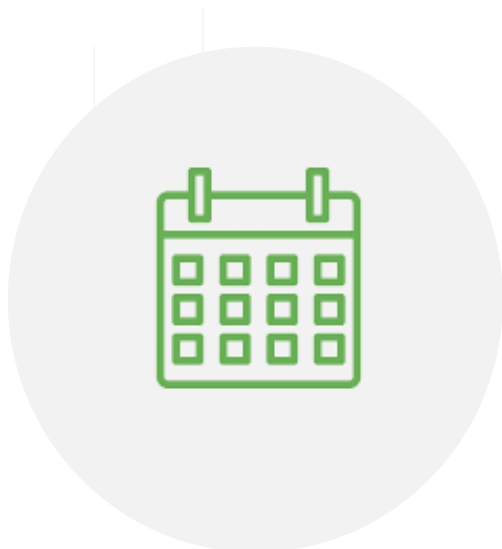
Quando os pontos do diagrama têm como imagem uma reta decrescente.

## Correlação linear nula

Quando os pontos do diagrama apresentam-se dispersos, não oferecendo imagem definida

# Dados de mandíbula de fetos - Mandible

Conjunto de dados contém 167 observações e 2 variáveis:



**age**

Idade gestacional em  
semanas



**length**

Comprimento da  
mandíbula em mm

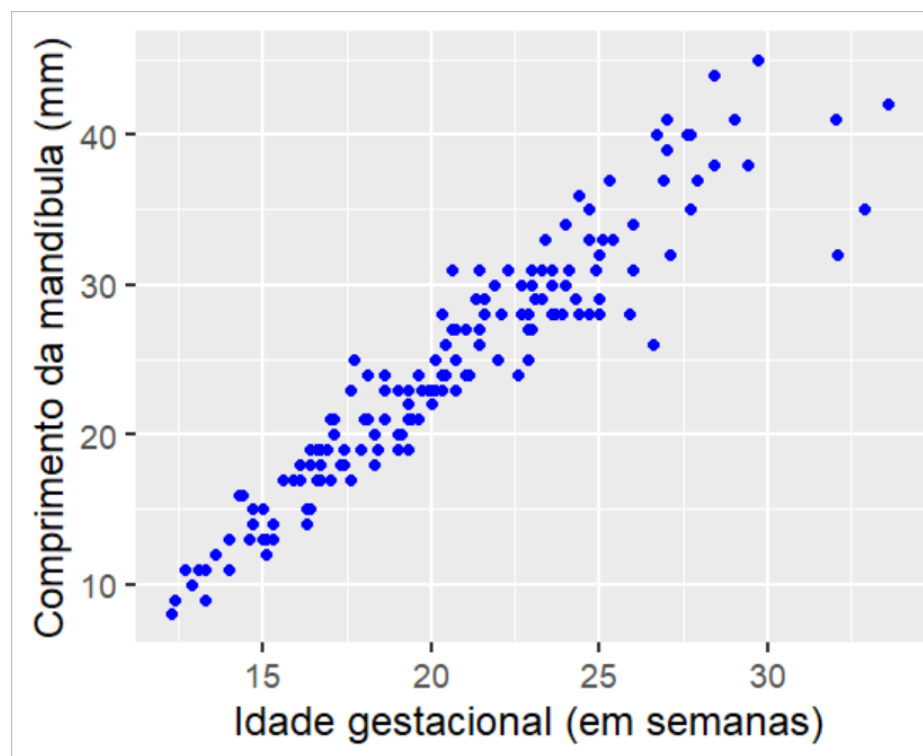
# Conjunto de dados “Mandible”



```
> head(lmtest::Mandible)
  age length
1 12.3     8
2 12.4     9
3 12.7    11
4 12.7    11
5 12.9    10
6 13.1    11
> tail(lmtest::Mandible)
  age length
162 29.4    38
163 29.7    45
164 32.0    41
165 32.1    32
166 32.9    35
167 33.6    42
```

# Mandible

*Com o aumento da idade gestacional, há um aumento no comprimento da mandíbula do feto?*



**Figura 2:** Correlação entre a idade gestacional e o comprimento da mandíbula do feto

# Coeficiente de correlação linear

Esse coeficiente deve indicar o grau de intensidade da correlação entre duas variáveis e, ainda, o sentido dessa correlação (positivo ou negativo).

O valor do coeficiente de correlação não deve depender da unidade de medida dos dados.

$$r = \frac{n \sum (x_i y_i) - (\sum x_i) (\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$



# Correlação age × length

Para calcular o coeficiente de correlação de Pearson para as variáveis age (X) e length (Y), vamos primeiramente construir uma tabela para os cálculos das variáveis.



	$x_i$	$y_i$	$x_i^2$	$y_i^2$	$x_i y_i$
1	12,3	8	151,29	1.444	98,4
2	12,4	9	153,76	2.025	111,6
3	12,7	11	161,29	2.025	139,7
4	12,7	11	161,29	1.681	139,7
5	12,9	10	166,41	1.024	129,0
⋮	⋮	⋮	⋮	⋮	⋮
167	33,6	42	1128,96	1.764	1411,2
Total	3.447,8	4.091	74.633,14	111.169	90.249,6

$$r = \frac{167 \times 90.249,6 - 3.447,8 \times 4.091}{\sqrt{167 \times 74.633,14 - (3.447,8)^2} \sqrt{167 \times 111.169 - (4.091)^2}}$$
$$= -\frac{966.733,4}{\sqrt{576.409,5} \sqrt{1.828.942}} = 0,9415452$$

Conforme era esperado, obtivemos para  $r$  um valor negativo e moderado, pois os pontos no diagrama de dispersão na [Figura 2](#) indicaram uma correlação linear moderada.

# Coeficiente de correlação linear - interpretação

Para qualquer conjunto de dados o valor do coeficiente de correlação de **Pearson**,  $r$ , estará no intervalo de  $-1$  a  $1$ . Será positivo quando os dados apresentarem correlação linear positiva; será negativo quando os dados apresentarem correlação linear negativa.

O valor de  $r$  será tão mais próximo de  $1$  (ou  $-1$ ) quando mais forte for a correlação dos dados observados. Teremos  $r = +1$  se os pontos estiverem exatamente sobre uma reta ascendente (correlação positiva perfeita). Por outro lado, teremos  $r = -1$  se os pontos estiverem exatamente sobre uma reta descendente (correlação negativa perfeita). Quando não houver correlação nos dados,  $r$  acusará um valor próximo de  $0$  (zero).

Correlação muito forte	Correlação forte	Correlação moderada	Correlação fraca	Correlação desprezível
$ 0,9  \leq r \leq  1,0 $	$ 0,7  \leq r <  0,9 $	$ 0,5  \leq r <  0,7 $	$ 0,3  \leq r <  0,5 $	$0 \leq r <  0,3 $

# Análise de regressão

Além de determinar se existe uma relação linear entre duas variáveis,  $X$  e  $Y$ , frequentemente se deseja conhecer a função que mostra como  $Y$  varia em função de  $X$ . Sempre que desejamos estudar determinada variável em função de outra, fazemos uma *análise de regressão*.

*Ao avaliar o comprimento da mandíbula e a idade gestacional do feto, pergunta-se:*

*com o passar do tempo, o comprimento da mandíbula do feto aumenta?*

# Regressão linear simples

Um estudo de regressão com a formulação mais simples relaciona uma variável  $Y$ , chamada de *resposta*, com uma variável  $X$ , denominada de variável *regressora*.

Assim, supondo  $X$  a variável regressora e  $Y$  a resposta, vamos procurar determinar o ajustamento de uma reta à relação entre essas variáveis. Definimos, então, o *modelo de regressão linear simples*, dado por:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Em que  $\beta_0$  é o *coeficiente linear*,  $\beta_1$  é o *coeficiente angular* e  $\varepsilon$  é um erro aleatório com média zero e variância  $\sigma^2$ .

$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\beta_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n}$$

$n$  é o número de observações

# Regressão linear simples

Como geralmente fazemos uso de uma amostra para obter os valores desses coeficientes, o resultado, na realidade, é uma estimativa da verdadeira equação de regressão. Sendo assim, escrevemos:

$$\hat{Y} = \beta_0 + \beta_1 X$$

$\hat{Y}$  é o  $Y$  estimado

## Erro aleatório $\varepsilon$

O erro aleatório  $\varepsilon$ , também chamado de *resíduo*, é dado pela diferença entre o valor observado e o respectivo valor estimado de  $Y$ , isto é,

$$\varepsilon = Y - \hat{Y}$$

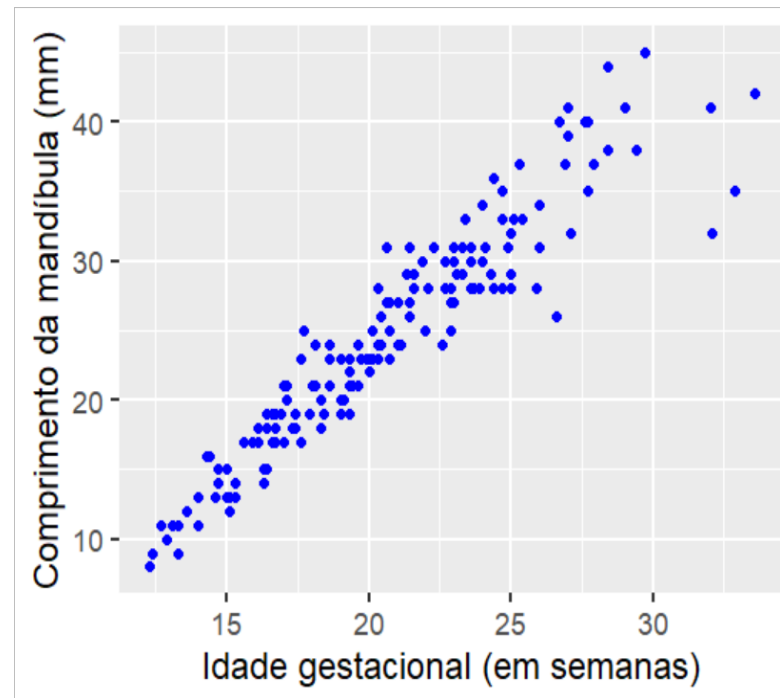
*Com o passar do tempo, o comprimento da mandíbula do feto aumenta?*



# Correlação age × length

Já sabemos que existe uma correlação positiva muito forte entre as variáveis ( $r = 0,94$ ): à medida que passam as semanas, maior o comprimento da mandíbula do feto.

	$x_i$	$y_i$
1	12,3	8
2	12,4	9
3	12,7	11
4	12,7	11
5	12,9	10
⋮	⋮	⋮
167	33,6	42



**Figura 2:** Correlação entre a idade gestacional e o comprimento da mandíbula do feto

# Equação de regressão age × length

Para a encontrar a equação de regressão linear é preciso calcular os coeficientes  $\beta_0$  e  $\beta_1$ .

	$x_i$	$y_i$	$x_i^2$	$x_i y_i$
1	12,3	8	151,29	98,4
2	12,4	9	153,76	111,6
3	12,7	11	161,29	139,7
4	12,7	11	161,29	139,7
5	12,9	10	166,41	129,0
⋮	⋮	⋮	⋮	⋮
167	33,6	42	1128,96	1411,2
Total	3.447,8	4.091	74.633,14	90.249,6

$$\beta_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} ; \beta_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n}$$

$$\beta_1 = \frac{167 \times 90.249,6 - 3.447,8 \times 4.091}{167 \times 74.633,14 - (3.447,8)^2} \cong 1,67$$

$$\beta_0 = \frac{4.091}{167} - 2,76 \times \frac{3.447,8}{167} = -10,1289$$

Assim, temos a seguinte equação de regressão:

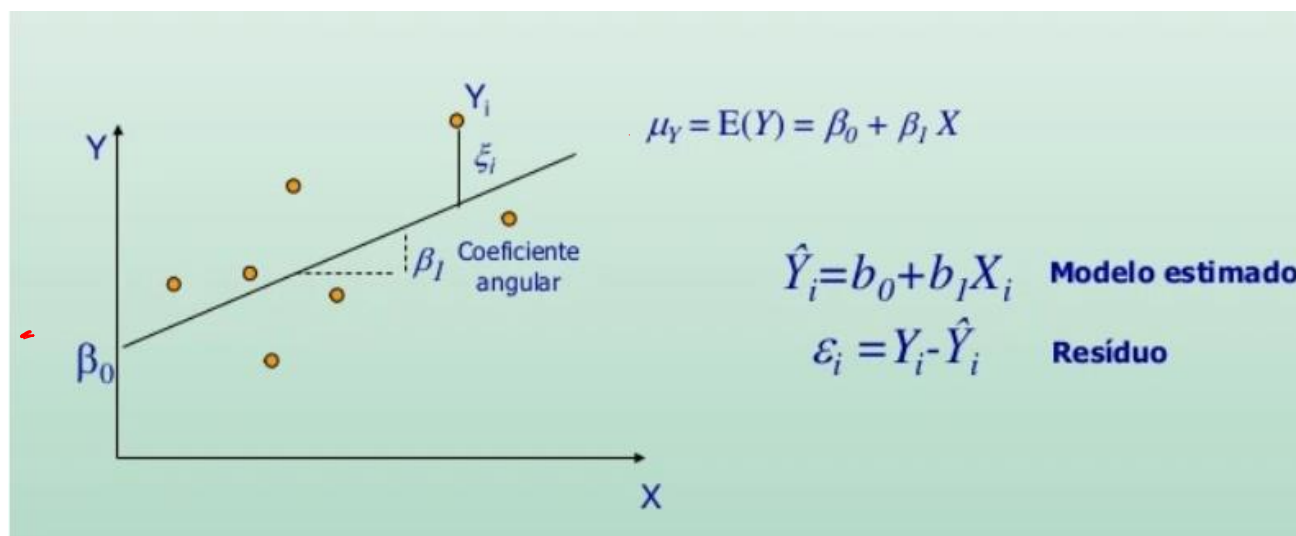
$$\hat{y} = \beta_0 + \beta_1 x$$

$$\hat{y} = -10,1289 + 1,67x$$



# A reta de regressão

Com a equação de regressão podemos aproximar por uma linha reta um determinado padrão, ou conjunto, de pontos. De regra não podemos traçar uma reta que passe por todos os pontos, mas podemos determinar uma reta que passe perto da maioria deles. Esse tipo de reta é chamada *reta de regressão* ou *reta de mínimos quadrados*.



# A reta de regressão

Para traçar a reta de regressão, basta atribuir dois valores para  $x$  e calcular os correspondentes valores de  $\hat{y}$ .

O menor valor para age, idade gestacional, 12,3 semanas e a maior é 33,6 semanas.

$$\hat{y} = -10,1289 + 1,67x$$

$$\hat{y} = -10,1289 + 1,67 \times 12,33$$

$$\hat{y}_{x=12,3} = 10,5$$

$$\hat{y} = -10,1289 + 1,67 \times 33,6$$

$$\hat{y}_{x=33,6} = 46,2$$

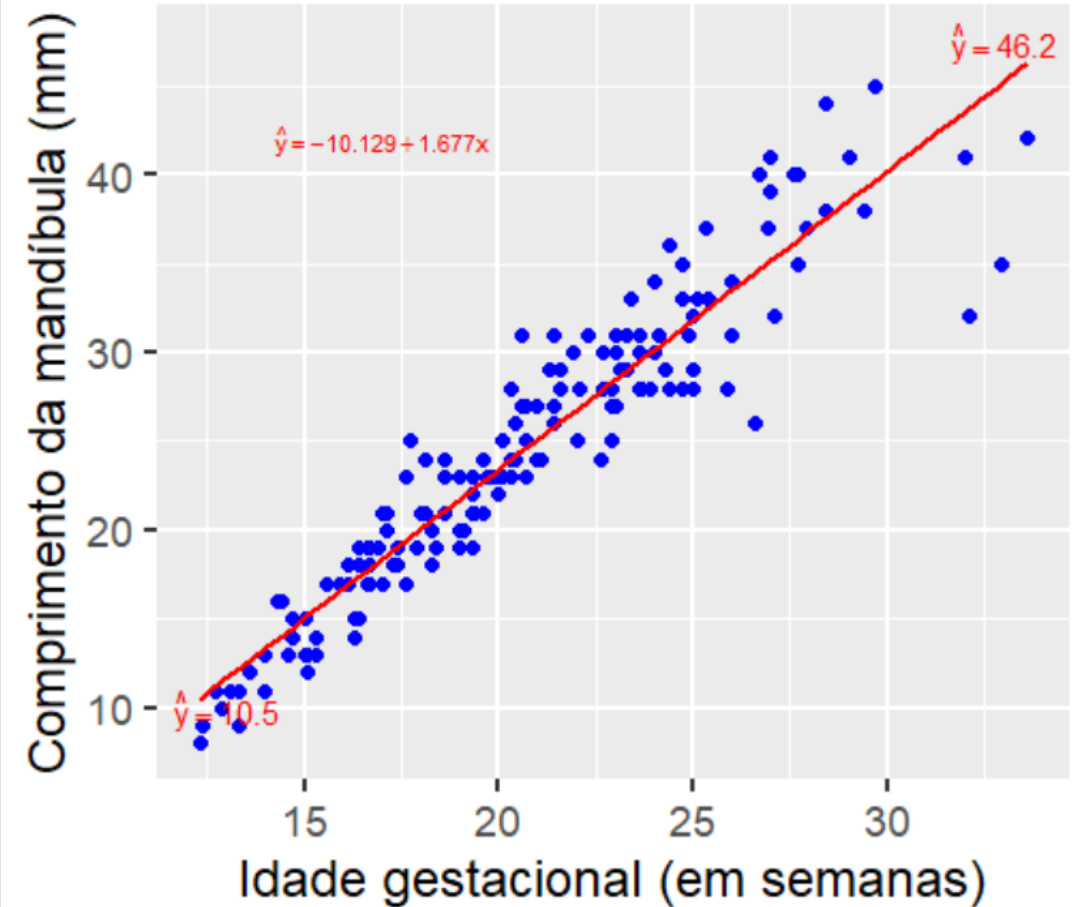


# A reta de regressão

Para traçar a reta de regressão, basta atribuir dois valores para  $x$  e calcular os correspondentes valores de  $\hat{y}$ .

$$x = 12,3 \Rightarrow \hat{y} = -10,1289 + 1,67 \times 12,3 = 10,5$$

$$x = 33,6 \Rightarrow \hat{y} = -10,1289 + 1,67 \times 33,6 = 46,2$$



**Figura 3:** Correlação entre a idade gestacional e o comprimento da mandíbula do feto e reta de regressão

# Estimativas através da equação de regressão

A partir dos ensaios experimentais, construímos um modelo, o qual nos permite prever o comprimento da mandíbula do feto,  $\text{length } (\hat{y})$ , a partir de uma determinada semana gestacional,  $\text{age } (x)$ . Por exemplo, se um feto está na décima terceira semana gestacional, esperamos comprimento da mandíbula igual a  $\hat{y} = -10,1289 + 1,67 \times 12 \cong 9,997$  e  $\hat{y} = -10,1289 + 1,67 \times 13 \cong 11,67$  para a terceira semana.

O coeficiente  $\beta_1$  fornece uma estimativa da variação esperada de  $Y$ , a partir da variação de uma unidade em  $X$ . O sinal deste coeficiente indica o sentido da variação. No exemplo, podemos dizer: a cada semana a mais na idade gestacional do feto, esperamos um aumento de 1,67 no comprimento da mandíbula.

Podemos confirmar fazendo a diferença para a estimativa do comprimento médio da mandíbula para a décima terceira e décima segunda semanas, por exemplo:  $11,67 - 9,997 = 1,67$ .

*A equação encontrada  
oferece um bom ajuste?*



$$\hat{y} = -10,289 + 1,67x$$

# Coeficiente de determinação



O coeficiente de determinação é a quantidade de variação em  $y$  que é explicada pela reta de regressão

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

$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  é a variação explicada e  $\sum_{i=1}^n (y_i - \bar{y})^2$  é a variação total.

Podemos calcular  $r^2$  usando a definição que acabamos de dar, ou podemos simplesmente elevar ao quadrado o coeficiente de correlação linear de Pearson.

Notamos que  $r = 0,9415452$ , então  $r^2 \cong 0,8865$ , o que significa que 88,65% da variação total do comprimento da mandíbula pode ser explicada pela variação da idade gestacional do feto.

Outra interpretação para o coeficiente de determinação é que ele mede a precisão da reta de regressão, assim, quanto maior o valor de  $r^2$ , melhor será o ajuste da reta aos dados.

# *Como melhorar o modelo de regressão?*



**Água**



**Cultivar**



**Nutrientes**

Se tivéssemos variáveis relacionadas a produção agrícola, com  $r^2 = 38,26\%$  , por exemplo, então cerca de 62% da variabilidade de y não pode ser descrita (ou explicada) pela variabilidade de x. Fica portanto claro que existem outros fatores que poderiam ser importantes, como por exemplo, quantidade de água recebida, espécie, tipos de nutrientes, etc.



# Youtube



## **Coeficiente de correlação**

<https://youtu.be/BxDaOWdRBqk>



## **Correlação e regressão**

[https://youtu.be/uF78\\_zMorHU](https://youtu.be/uF78_zMorHU)



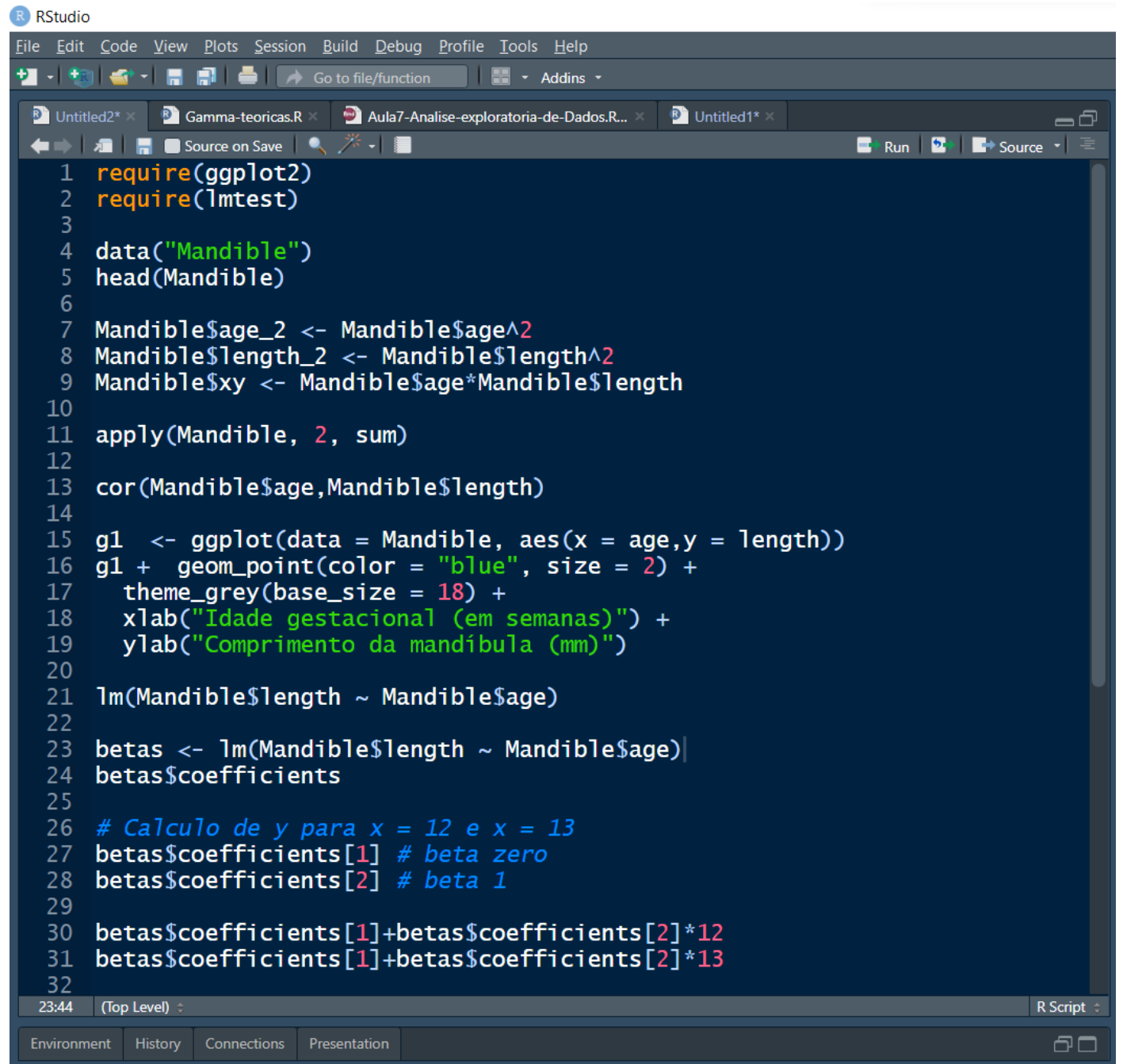
## **Coeficiente de determinação**

<https://youtu.be/L0kBCq1Qmbk>



# Em R

- Organização dos dados
- Análise de correlação
- Análise de regressão



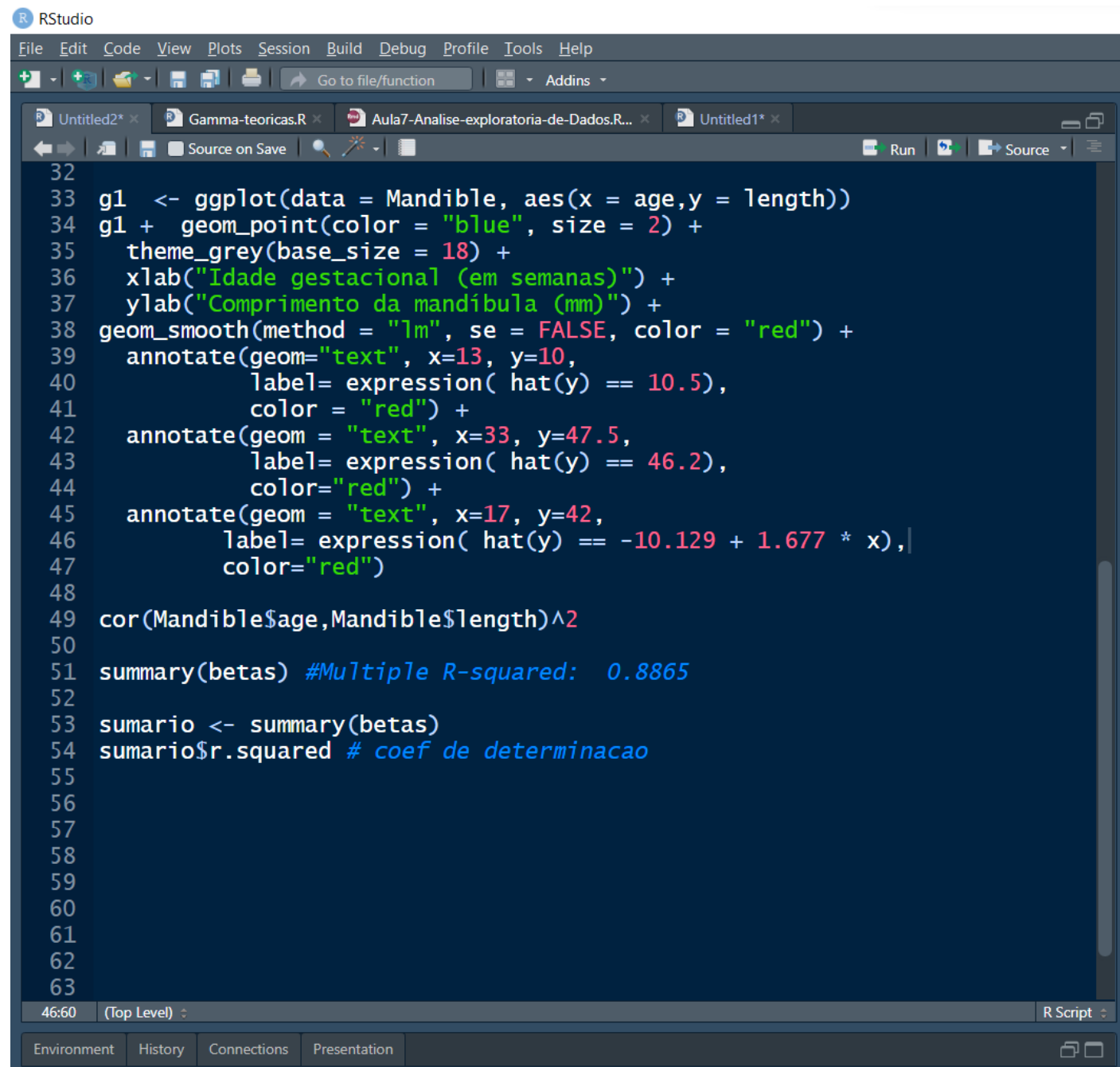
The screenshot shows the RStudio environment with a script editor containing the following R code:

```
1 require(ggplot2)
2 require(lmtest)
3
4 data("Mandible")
5 head(Mandible)
6
7 Mandible$age_2 <- Mandible$age^2
8 Mandible$length_2 <- Mandible$length^2
9 Mandible$xy <- Mandible$age*Mandible$length
10
11 apply(Mandible, 2, sum)
12
13 cor(Mandible$age,Mandible$length)
14
15 g1 <- ggplot(data = Mandible, aes(x = age,y = length))
16 g1 + geom_point(color = "blue", size = 2) +
17   theme_grey(base_size = 18) +
18   xlab("Idade gestacional (em semanas)") +
19   ylab("Comprimento da mandíbula (mm)")
20
21 lm(Mandible$length ~ Mandible$age)
22
23 betas <- lm(Mandible$length ~ Mandible$age)
24 betas$coefficients
25
26 # Calculo de y para x = 12 e x = 13
27 betas$coefficients[1] # beta zero
28 betas$coefficients[2] # beta 1
29
30 betas$coefficients[1]+betas$coefficients[2]*12
31 betas$coefficients[1]+betas$coefficients[2]*13
32
```

The RStudio interface includes a menu bar (File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help), a toolbar with icons for file operations and running code, and a tab bar showing open files: 'Untitled2\*', 'Gamma-teoricas.R', 'Aula7-Analise-exploratoria-de-Dados.R...', and 'Untitled1\*'. The status bar at the bottom shows '23:44', '(Top Level)', and 'R Script'.

# Em R

- Gráfico de correlação com reta de regressão
- Coeficiente de correlação
- Coeficiente de determinação



```
32
33 g1 <- ggplot(data = Mandible, aes(x = age, y = length))
34 g1 + geom_point(color = "blue", size = 2) +
35   theme_grey(base_size = 18) +
36   xlab("Idade gestacional (em semanas)") +
37   ylab("Comprimento da mandíbula (mm)") +
38   geom_smooth(method = "lm", se = FALSE, color = "red") +
39   annotate(geom="text", x=13, y=10,
40           label= expression( hat(y) == 10.5),
41           color = "red") +
42   annotate(geom = "text", x=33, y=47.5,
43           label= expression( hat(y) == 46.2),
44           color="red") +
45   annotate(geom = "text", x=17, y=42,
46           label= expression( hat(y) == -10.129 + 1.677 * x),
47           color="red")
48
49 cor(Mandible$age, Mandible$length)^2
50
51 summary(betas) #Multiple R-squared:  0.8865
52
53 sumario <- summary(betas)
54 sumario$r.squared # coef de determinacao
55
56
57
58
59
60
61
62
63
```

46:60 (Top Level) R Script

Environment History Connections Presentation