Introdução à estatística

Aula 10 - Introdução à Regressão Simples

Felipe Nunes, Ph.D.

November 20, 2016

Curso de aperfeiçoamento BR040

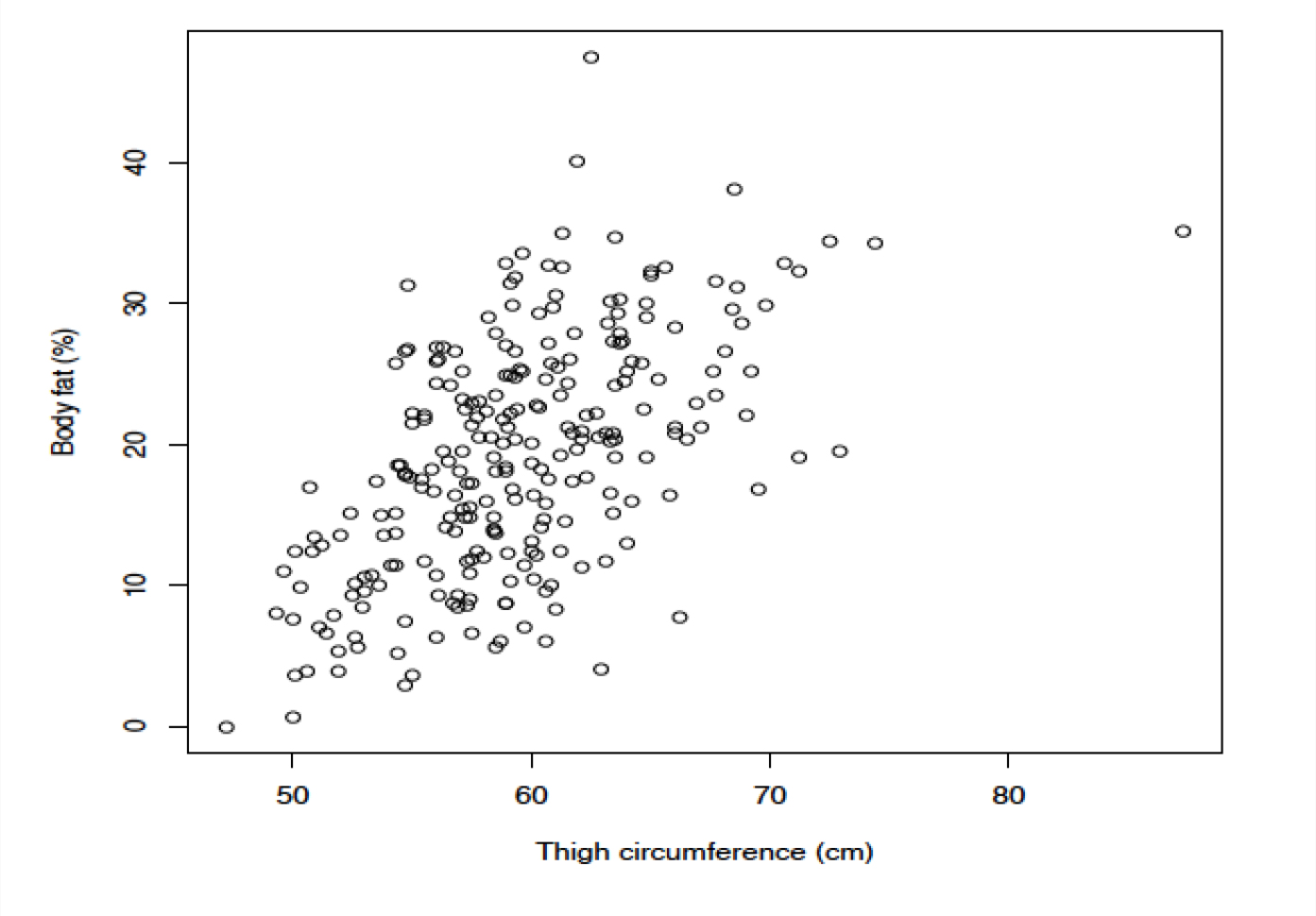
Introdução

Correlação

- Análise de regressão é um método estatístico que busca descobrir como uma variável está **relacionada** à outra.
- Esta técnica é útil por ser capaz de prever valores de uma variável utilizando outras.
- Vamos ver um exemplo para que fique mais claro como análises multivariadas desse tipo funcionam.

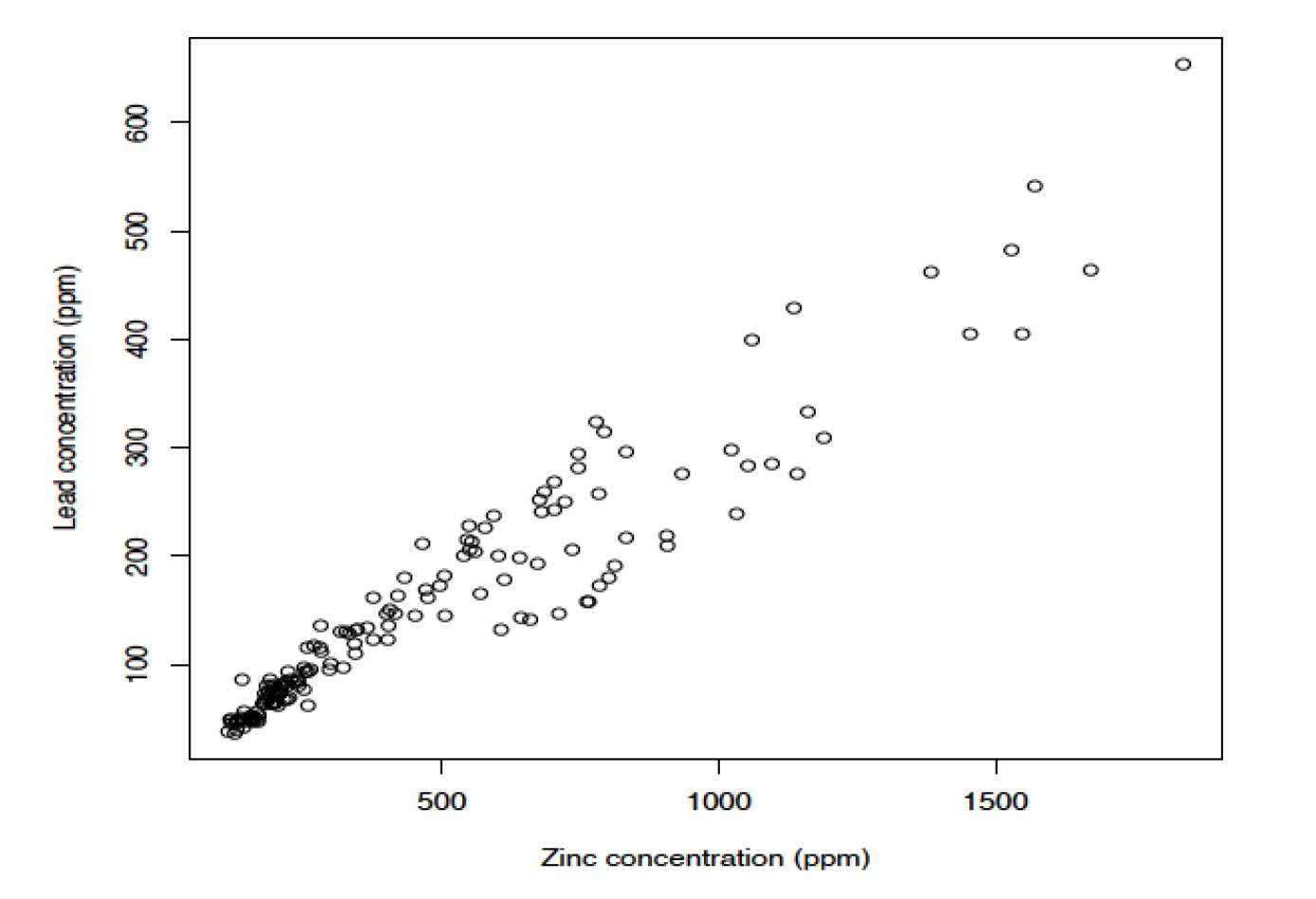
Correlação

Percentual de gordura no corpo contra a dimensão da circunferência abdominal (cm).



Correlação

Concentração de chumbo contra a concentração de zinco em água.



Regressão

- O que estamos vendo nas figuras anteriores?
- Há alguma equação que pode modelar a associação entre as variáveis nas figuras mostradas?

Regressão

- Equação do modelo de regressão linear simples:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon$$

$$\epsilon \sim N(0, \sigma)$$

  - y = variável dependente (variável aleatória)
  - x = variável independente (não aleatória)
  - β_0 = intercepto (não aleatória)
  - β_1 = inclinação (não aleatória)
  - ϵ = erro da estimativa (aleatório)

Regressão

- Como calcular (estimar) os valores de β_0 e β_1?
- Vamos usar o **método de mínimos quadrados ordinários**
- Para o uso desse método é necessário que haja algum convencimento a respeito do comportamento das variáveis e da relação entre elas.

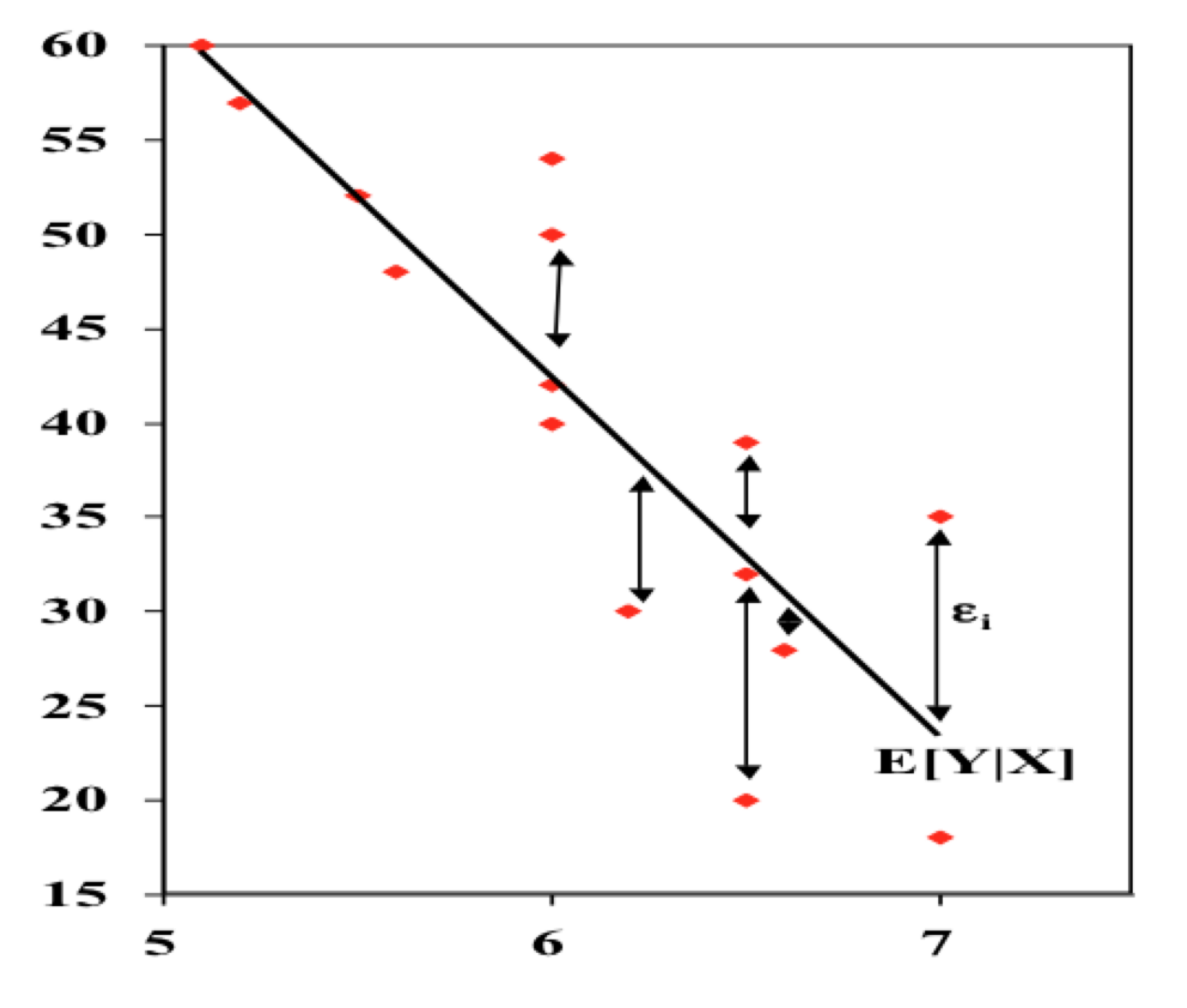
Regressão

- Pressupostos (hipóteses básicas):
  - Relacionamento linear entre as variáveis
  - $\mathrm{E}(\epsilon) = 0$
  - $\mathrm{E}(\epsilon^2) = \sigma^2$ (constante)
  - Os resíduos são independentes entre si: $E(\epsilon_i, \epsilon_j) = 0$
  - Os resíduos e as variáveis são independentes: $E(x, \epsilon) = 0$
  - As variáveis x_n não podem ser combinações lineares entre si

Regressão

- O ajuste da regressão: graficamente, a análise de regressão implica no ajuste de uma reta que represente uma 'boa forma' da estrutura dos dados.
- Mas o que é 'boa forma' de ajuste da reta?
    - Vamos observar a diferença entre a reta ajustada (que é produto do valor esperado condicional) e a observação realizada correspondente ao resíduo.
    - Logo, o ajuste ideal da reta deve respeitar a condição de 'menor distância possível' em relação aos valores observados.

Regressão



Regressão

- Logo, a idéia de ajuste dos parâmetros do valor esperado condicional passa por 'Minimizar a Soma dos Quadrados dos Resíduos".
- O estimador de Mínimos Quadrados Ordinários possui propriedades interessantes, quando as hipóteses básicas não são violadas: ele é **não-viesado** e é o mais **eficiente** entre os estimadores lineares.

Estimativa dos parâmetros com MQO

Estimativa dos parâmetros com MQO

- Usando MQO temos $\hat{\beta}_0$ e $\hat{\beta}_1$:

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

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

- A linha de regressão ficaria assim, então:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Estimativa dos parâmetros com MQO

- A distribuição dos parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$ poderia ser definida:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}\right)$$

Estimativa dos parâmetros com MQO

- O desvio-padrão é desconhecido e é estimado pelo 'erro padrão dos resíduos' - que mede a variabilidade em volta da linha de regressão estimada.
- O 'erro padrão dos resíduos' é computado por:

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

sendo que $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ é chamado de 'resíduo' da regressão por representar a diferença entre o valor observado e o valor estimado da regressão.

Estimativa dos parâmetros com MQO

- Outra estimativa importante é a do **coeficiente de determinação** da regressão (representado por R^2).
- Por meio desse indicador é possível encontrar o quão ajustada está a reta de regressão aos dados coletados.
- A variação total de y (soma dos quadrados totais) pode ser encontrada somando a soma dos quadrados da regressão com a soma dos quadrados do erro padrão da regressão:

$$\begin{aligned} SQT &= SQR + SQE \\ \sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \end{aligned}$$

Estimativa dos parâmetros com MQO

- O percentual da variação de y que pode ser explicada por x (coeficiente de determinação), pode ser obtido por:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}$$

  sendo que $0 \leq R^2 \leq 1$

- Também pode-se representar:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Estimativa dos parâmetros com MQO

- O coeficiente de correlação pode ser encontrado assim:

$$r = \hat{\beta}_1 \frac{s_x}{s_y}$$

  sendo que s_x e s_y são os desvios-padrão de x e y

- E a covariança entre y e x:

$$cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Estimativa dos parâmetros com MQO

- O erro padrão de $\hat{\beta}_1$ e $\hat{\beta}_0$:

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$s_{\hat{\beta}_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

Estimativa dos parâmetros com MQO

- Para testar a relação linear entre y e x formulamos um teste de hipótese:

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

- A estatística de teste:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}$$

  Rejeita-se a hipótese nula se $t > t_{\alpha/2, n-2}$

Estimativa dos parâmetros com MQO

- Intervalo de confiança para β_1:

$$\hat{\beta}_1 - t_{\alpha/2,n-2} \times s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2} \times s_{\hat{\beta}_1}$$

Exemplo

Exemplo

Os dados abaixo mostram a distância percorrida por um carro (y) pelo volume de octanagem da gasolina para teste (x):

	y	x	xy	y^2	x^2
	13	89	1157	169	7921
	13.5	93	1255.5	182.25	8649
	13	87	1131	169	7569
	13.2	90	1188	174.24	8100
	13.3	89	1183.7	176.89	7921
	13.8	95	1311	190.44	9025
	14.3	100	1430	204.49	10000
	14.0	98	1372	196	9604
Soma	108.1	741	10028.2	1462.31	68789

Exemplo

Então já temos...

$$\sum y_i = 108.1$$

$$\sum x_i = 741$$

$$\sum x_i y_i = 10028.2$$

$$\sum y_i^2 = 1462.31$$

$$\sum x_i^2 = 68789$$

Exemplo

- Para encontrar a estimativa de $\hat{\beta}_1$ usando MQO:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum x_i y_i - 1/n(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \\
&= \frac{10028.2 - 1/8(741)(108.1)}{68789 - 741^2/8} \\
&= 0.10
\end{aligned}$$

- Para encontrar a estimativa de $\hat{\beta}_0$:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= \frac{108.1}{8} - 0.10\frac{741}{8} \\
&= 4.22
\end{aligned}$$

Exemplo

- Usando os valores estimados é possível encontrar valores preditos para cada valor de x:

$$\hat{y}_i = 4.22 + 0.10(x_i)$$

- Por exemplo, o valor predito para o primeiro caso do banco de dados é:

$$\hat{y}_i = 4.22 + 0.10(89) = 13.15$$

e o valor residual para o primeiro caso:

$$e_1 = y_1 - \hat{y}_1 = 13 - 13.15 = -0.15$$

Exemplo

A tabela completa nos dá:

$\hat{y}_i$	e_i	e_i^2
13.14883	-0.14882	0.02215
13.55013	-0.05013	0.00251
12.94818	0.05183	0.00269
13.24915	-0.04915	0.00242
13.14883	0.15118	0.02285
13.75078	0.04922	0.00242
14.25240	0.04760	0.00227
14.05175	-0.05175	0.00268
	$\sum e_i = 0$	$\sum e_i^2 = 0.06$

Exemplo

- Para estimar σ^2:

$$s_e^2 = \frac{\sum e_i^2}{n-2} = \frac{0.06}{8-2} = 0.0099$$

- O que nos dá:

$$s_e = \sqrt{0.0099} = 0.099$$

- Agora podemos calcular o erro padrão de $\hat{\beta}_1$:

$$s_{\hat{\beta}_1} = \frac{s_e}{\sqrt{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}} = \frac{0.099}{\sqrt{68789 - \frac{741^2}{8}}} = 0.00806$$

Exemplo

- O intervalo de confiança de 95% para β_1:

$$Pr(\hat{\beta}_1 - t_{\alpha/2,n-2} \times s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2} \times s_{\hat{\beta}_1}) = 1 - \alpha$$

$$0.10 - 2.447(0.00806) \leq \beta_1 \leq 0.10 + 2.447(0.00806)$$

$$0.08 \leq \beta_1 \leq 0.12$$

- De forma que nós estamos 95% confiantes que β_1 está entre 0.08 e 0.12.

Exemplo

- Se quisermos estimar o consumo de gasolina para uma gasolina com octanagem igual a 94:

$$\hat{y} = 4.22 + 0.10(94) = 13.65$$

- E podemos definir o coeficiente de determinação desse modelo:

$$R^2 = 1 - \frac{SQR}{SQT} = 1 - \frac{\sum e_i^2}{(n-1)s_y^2} = 1 - \frac{0.06}{7 \times (0.23)} = 0.96$$

Portanto, 96% da variação de y pode ser explicada por x. Trata-se de uma associação muito forte.

E no R? É mais fáçil?

Exemplo no R

```
#Enter the data:
x <- c(89,93,87,90,89,95,100,98)
y <- c(13,13.5,13,13.2,13.3,13.8,14.3,14)

#Run the regression of y on x:
ex <- lm(y~x)

#Display the results:
summary(ex)
```

Exemplo no R

```
Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-0.148822 -0.050528 -0.000772  0.049878  0.151178

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.21990    0.74743   5.646  0.00132 **
x            0.10032    0.00806  12.447 1.64e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09999 on 6 degrees of freedom
Multiple R-squared:  0.9627,    Adjusted R-squared:  0.9565
F-statistic: 154.9 on 1 and 6 DF,  p-value: 1.643e-05
```