Modelos Lineares Generalizados - Introdução a Regresssão Linear

Antonio C. da Silva Júnior

05 de dezembro de 2020

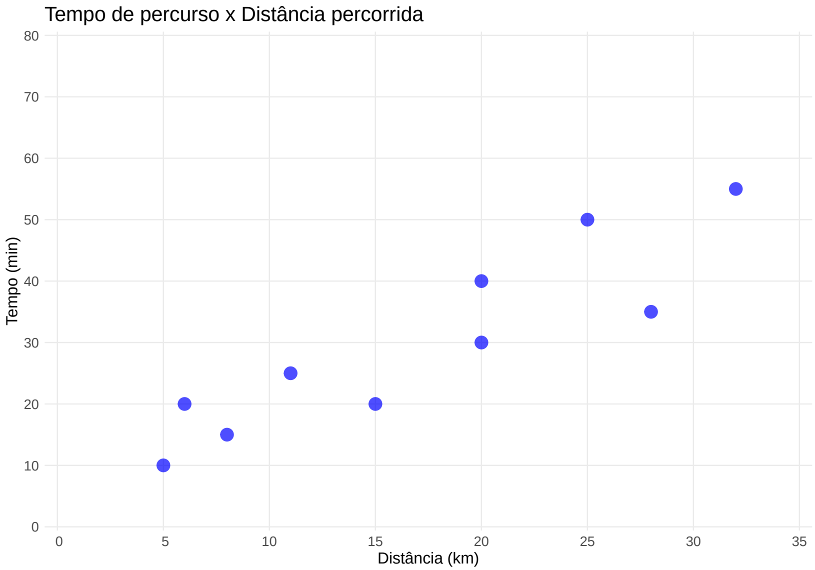
Conceitos gerais

Exemplo

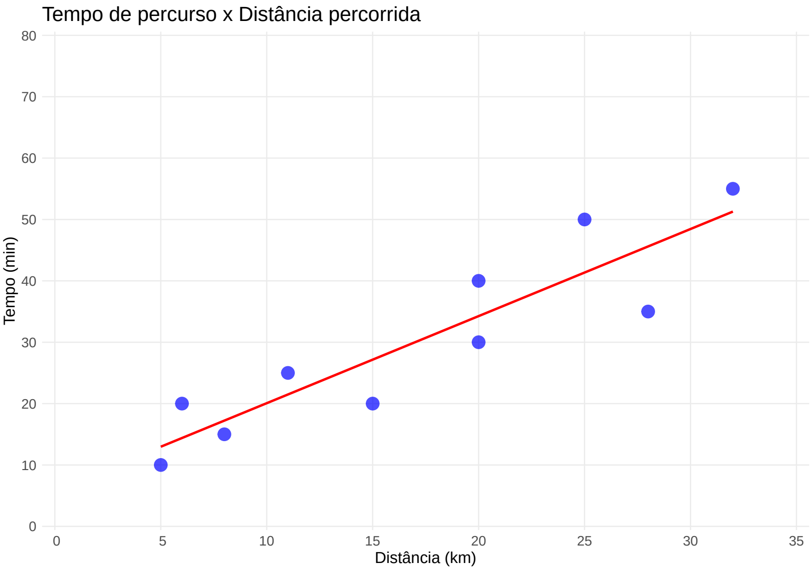
Estudar o comportamento do "tempo de percurso até a escola" em função da "distância percorrida pelos alunos".

Tempo (min)	Distância (km)
15	8
20	6
20	15
40	20
50	25
25	11
10	5
55	32
35	28
30	20

Exemplo


Tempo de percurso x Distância percorrida
Tempo (min)
0
10
20
30
40
50
60
70
80
0
5
10
15
20
25
30
35
Distância (km)

Exemplo


Tempo de percurso x Distância percorrida
Tempo (min)
80
70
60
50
40
30
20
10
0
0
5
10
15
20
25
30
35
Distância (km)

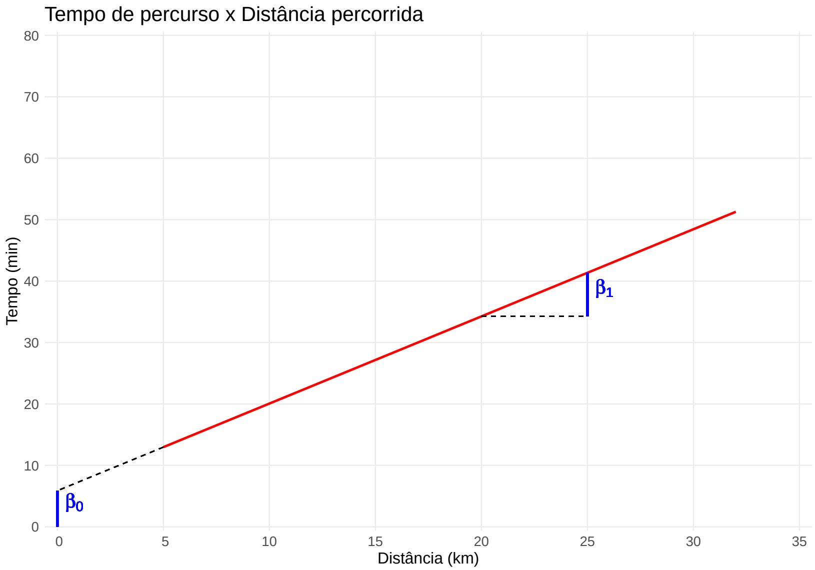
Regressão Linear Simples

- Equação do modelo:

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{1}$$

em que β_0 é o intercepto (coeficiente linear), β_1 é a inclinação da reta (coeficiente angular) e ϵ, o termo de erro aleatório.

Parâmetros


Tempo de percurso x Distância percorrida
80
70
60
50
40
30
20
10
0
Tempo (min)
0
5
10
15
20
25
30
35
Distância (km)
β_1
β_0

Estimativas

- Modelo ajustado:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{2}$$

com $\hat{\beta}_0 = 5.8784$ e $\hat{\beta}_1 = 1.4189$.

Estimativas

Portanto, um aluno que percorre uma distância de 25km, leva em média 41,3 minutos para chegar na escola.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 25$$

$$41.3509 = 5.8784 + 1.4189 \times 25$$

Interpretação

A cada k unidades a mais de distância percorrida, o tempo de um aluno chegar na escola aumenta, em média, $k\hat{\beta}_1$ unidades de tempo.

Dentro do contexto, 10km a mais de distância percorrida aumenta, em média, 14,19 minutos $(10 \times 1,4189)$ o tempo do trajeto.

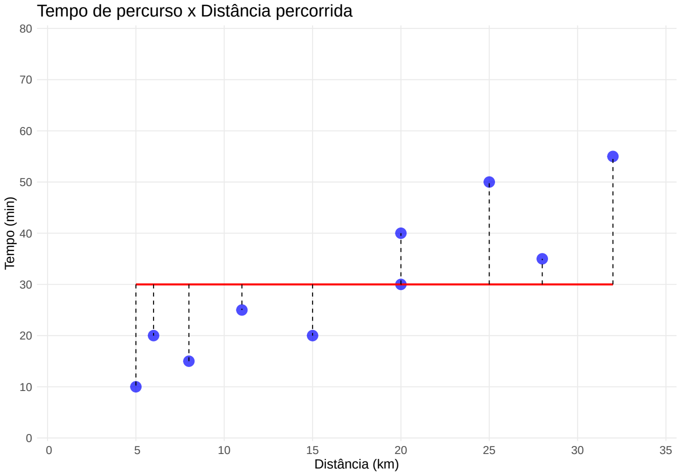
Interpretação

E caso a distância percorrida seja 0, o tempo para chegar na escola será de 5,8784 ($\hat{\beta}_0$) minutos.

Matematicamente a afirmação faz sentido, entretanto, ao observar que nos dados de ajuste do modelo não havia nenhum aluno com distância percorrida próxima de 0, concluimos que o intercepto indica, na verdade, uma extrapolação da reta de regressão.

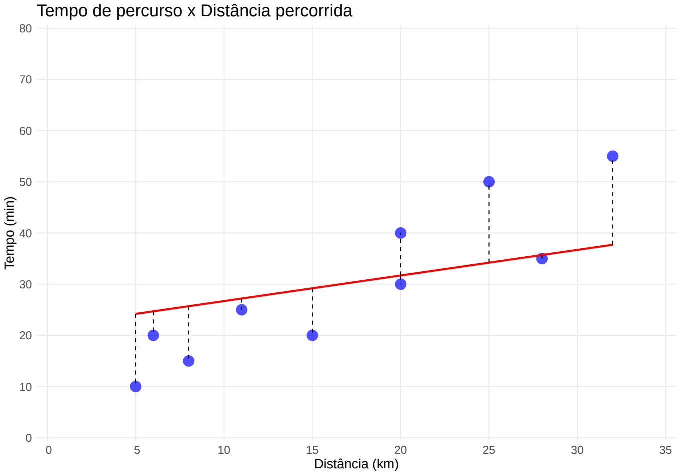
Ajuste do modelo

Estimação dos parâmetros



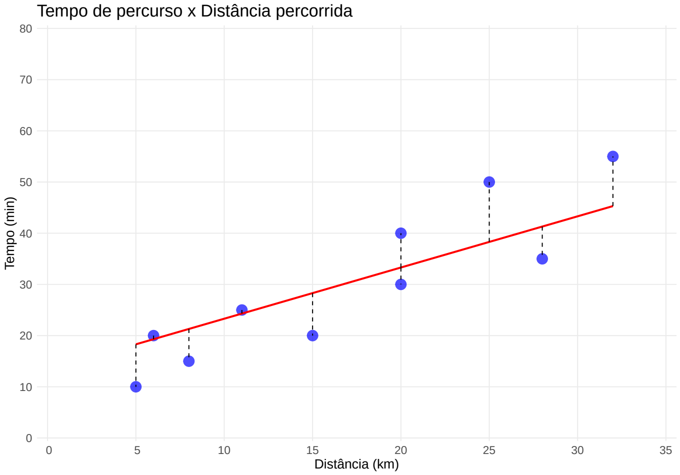
Reta	SQRes
0	2000

Estimação dos parâmetros



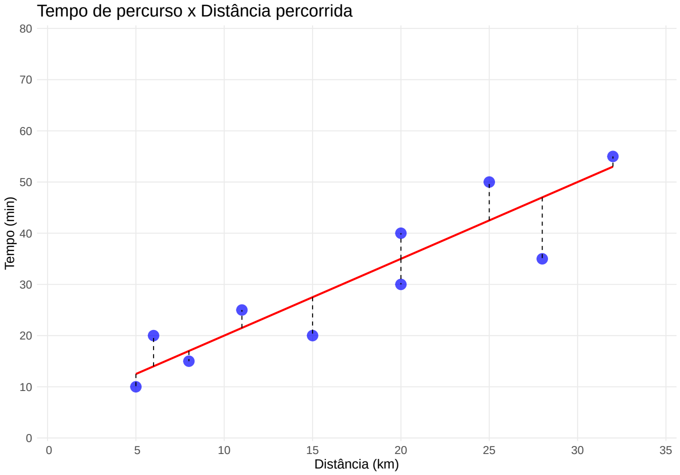
Reta	SQRes
0	2000.0
1	1048.9

Estimação dos parâmetros



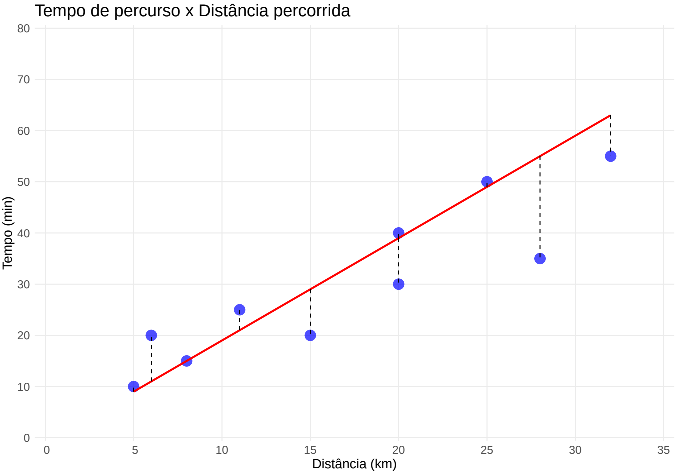
Reta	SQRes
0	2000.0
1	1048.9
2	504.9

Estimação dos parâmetros



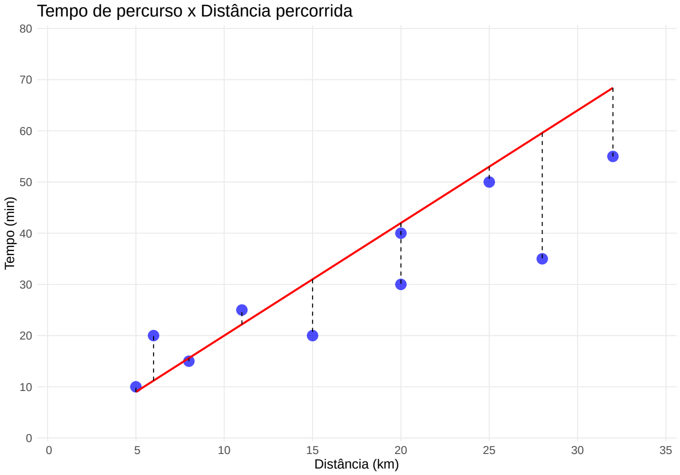
Reta	SQRes
0	2000.0
1	1048.9
2	504.9
3	369.0

Estimação dos parâmetros



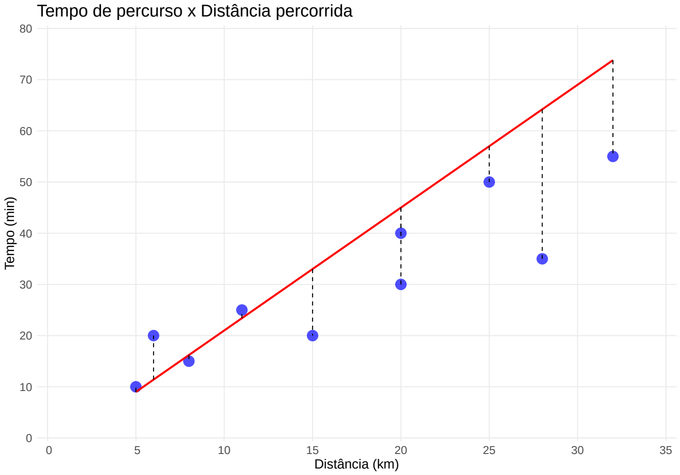
Reta	SQRes
0	2000.0
1	1048.9
2	504.9
3	369.0
4	726.0

Estimação dos parâmetros



Reta	SQRes
0	2000.00
1	1048.90
2	504.90
3	369.00
4	726.00
5	1149.36

Estimação dos parâmetros



Reta	SQRes
0	2000.00
1	1048.90
2	504.90
3	369.00
4	726.00
5	1149.36
6	1753.04

Estimação dos parâmetros

Estimação dos parâmetros

- Resíduo:

$$r_i = y_i - \hat{y}_i,$$

$$r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i),$$

$$r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Estimação dos parâmetros

- Quadrado do resíduo:

$$r_i^2 = (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- Soma dos quadrados dos resíduos:

$$S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{3}$$

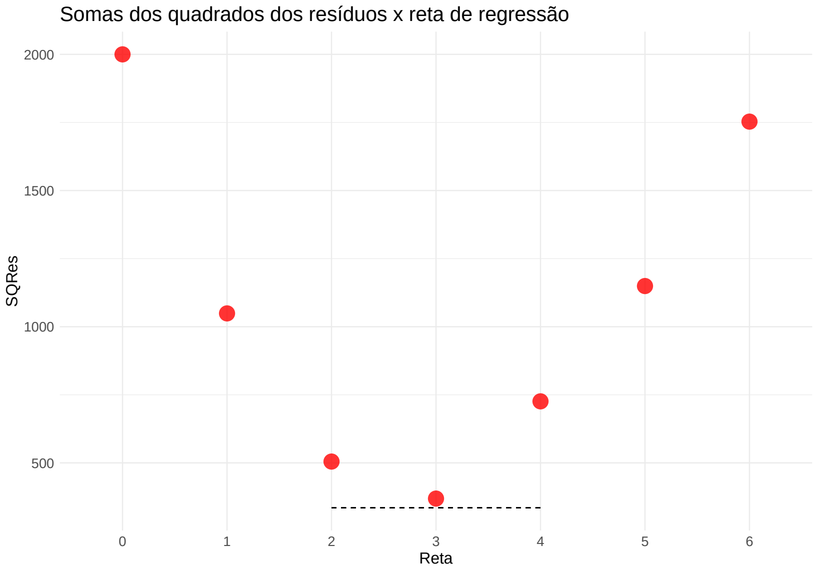
Estimação dos parâmetros

y (obs)	y (ajust)	Res	QRes
15	17.22973	-2.229730	4.971695
20	14.39189	5.608108	31.450877
20	27.16216	-7.162162	51.296567
40	34.25676	5.743243	32.984843
50	41.35135	8.648649	74.799123
25	21.48649	3.513514	12.344777
10	12.97297	-2.972973	8.838568
55	51.28378	3.716216	13.810263
35	45.60811	-10.608108	112.531958
30	34.25676	-4.256757	18.119978

Estimação dos parâmetros

Como estimar β_0 e β_1 de modo que a soma dos quadrados dos resíduos seja mínima?

Estimação dos parâmetros



Métodos dos Mínimos Quadrados Ordinários

- Derivada parcial em relação a $\hat{\beta}_0$:

$$\frac{\partial S}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

$$u = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

Métodos dos Mínimos Quadrados Ordinários

- Derivada parcial em relação a $\hat{\beta}_0$:

$$\frac{\partial S}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (u)^2,$$

$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} 2uu',$$

$$u' = -1,$$

Métodos dos Mínimos Quadrados Ordinários

- Derivada parcial em relação a $\hat{\beta}_0$:

$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1),$$

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i). \tag{4}$$

Métodos dos Mínimos Quadrados Ordinários

- Iguala a zero:

$$\frac{\partial S}{\partial \beta_0} = 0,$$

$$-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

$$-2\sum_{i=1}^{n} y_i + 2\sum_{i=1}^{n}\hat{\beta}_0 + 2\sum_{i=1}^{n}\hat{\beta}_1 x_i = 0,$$

Métodos dos Mínimos Quadrados Ordinários

- Iguala a zero:

$$\frac{-2\sum_{i=1}^{n} y_i}{2n} + \frac{2\sum_{i=1}^{n} \hat{\beta}_0}{2n} + \frac{2\sum_{i=1}^{n} \hat{\beta}_1 x_i}{2n} = \frac{0}{2n},$$

$$-\bar{y} + \hat{\beta}_0 + \hat{\beta}_0 \bar{x} = 0,$$

em que $\bar{y}$ e $\bar{x}$ são as médias amostrais de y e x, respectivamente.

Portanto,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{5}$$

Métodos dos Mínimos Quadrados Ordinários

- Derivada parcial em relação a $\hat{\beta}_1$:

$$\frac{\partial S}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

$$u = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

Métodos dos Mínimos Quadrados Ordinários

- Derivada parcial em relação a $\hat{\beta}_0$:

$$\frac{\partial S}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (u)^2,$$

$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} 2uu',$$

$$u' = -x_i,$$

Métodos dos Mínimos Quadrados Ordinários

- Derivada parcial em relação a $\hat{\beta}_0$:

$$\frac{\partial S}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i),$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i). \tag{6}$$

Métodos dos Mínimos Quadrados Ordinários

- Iguala a zero:

$$\frac{\partial S}{\partial \beta_1} = 0,$$

$$-2\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

,

Métodos dos Mínimos Quadrados Ordinários

Resolve o sistema substituindo $\hat{\beta}_0$ pela equação (5):

$$-2\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i) = 0,$$

$$-2\sum_{i=1}^{n} x_i(y_i - \bar{y} + \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i) = 0,$$

$$-2\sum_{i=1}^{n} x_i(y_i - \bar{y} + \hat{\beta}_1(\bar{x} - x_i)) = 0,$$

$$-2\sum_{i=1}^{n} x_i(y_i - \bar{y}) + \hat{\beta}_1 x_i(\bar{x} - x_i) = 0,$$

Métodos dos Mínimos Quadrados Ordinários

Resolve o sistema substituindo $\hat{\beta}_0$ pela equação (5):

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) + \hat{\beta}_1 x_i(\bar{x} - x_i) = 0,$$

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) + \hat{\beta}_1 \sum_{i=1}^{n} x_i(\bar{x} - x_i) = 0,$$

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

Métodos dos Mínimos Quadrados Ordinários

Resolve o sistema substituindo $\hat{\beta}_0$ pela equação (5):

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

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

Métodos dos Mínimos Quadrados Ordinários

Resolve o sistema substituindo $\hat{\beta}_0$ pela equação (5):

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

$$\hat{\beta}_1 = \frac{\bar{y} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i}{\bar{x} \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i^2}. \tag{7}$$

Ajuste do modelo

Exemplos no Excel. . .

Ajuste do modelo

Exemplo no R:

```
modelo <- lm(formula = Tempo ~ Distancia, data = df_tempo)
```

Ajuste do modelo

```r
summary(modelo)
```

```
##
## Call:
## lm(formula = Tempo ~ Distancia, data = df_tempo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6081  -3.9358   0.6419   5.1351   8.6486
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.8784     4.5323   1.297 0.230788
## Distancia     1.4189     0.2355   6.025 0.000314 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.719 on 8 degrees of freedom
## Multiple R-squared:  0.8194, Adjusted R-squared:  0.7969
## F-statistic:  36.3 on 1 and 8 DF,  p-value: 0.0003144
```