Avaliação 2

Metodos de interpolaçãoe mínimos quadrados

Paulo Ricardo Seganfredo Campana

13 de abril de 2023

Método de Mínimos quadrados

```r
library(tidyverse)
```

```r
min_quad <- function(dados, caso = "linear") {
    dados |>
        rename(x = 1, y = 2) |>
        mutate(y = case_when(
            caso == "linear"      ~      y,
            caso == "exponencial" ~  log(y),
            caso == "quadrático"  ~ sqrt(y)
        )) |>
        summarise(
            beta = sum((x - mean(x)) * (y - mean(y))) /
                   sum((x - mean(x))^2),
            alpha = mean(y) - beta * mean(x),
            desvio_total = case_when(
                caso == "linear"      ~ sum((    alpha + beta * x    -     y  )^2),
                caso == "exponencial" ~ sum((exp(alpha + beta * x)   - exp(y) )^2),
                caso == "quadrático"  ~ sum((   (alpha + beta * x)^2 -     y^2)^2)
            ),
            `R²` = cor(x, y)^2
        ) |>
        relocate(alpha, .before = beta)
}
```

A função `min_quad` calcula e resume os coeficientes da regressão linear $\hat{y} = \alpha + \beta x$, o desvio total D e o coeficiente de determinação R^2 de tal modo:

$$\beta = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$D = \sum_{i=1}^{n}(\alpha + \beta x_i - y_i)^2$$

$$\alpha = \bar{y} - \beta\bar{x}$$

$$R^2 = \frac{\left(\sum(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}$$

E para quando os dados se encaixam melhor em um modelo exponencial: $\alpha e^{\beta x}$ ou quadrático: $\alpha + \beta x^2$, a variável `caso` faz uma transformação inversa para que o método linear possa ser aplicado, a função `min_quad_fun` desfaz essa transformação para montar o modelo final, que será usado para intrapolar dados e gerar um gráfico do ajuste com `min_quad_plot`.

Geração de função e gráfico

```r
min_quad_fun <- function(dados, caso = "linear") {
    coef <- min_quad(dados, caso)
    function(x) case_when(
        caso == "linear"      ~     coef$alpha + coef$beta * x,
        caso == "exponencial" ~ exp(coef$alpha + coef$beta * x),
        caso == "quadrático" ~     (coef$alpha + coef$beta * x)^2
    )
}
```

```r
min_quad_plot <- function(dados, caso = "linear") {
    f <- min_quad_fun(dados, caso)
    dados |>
        rename(x = 1, y = 2) |>
        ggplot(aes(x, y)) +
        geom_function(fun = f, color = "#00c060") +
        geom_point(color = "#008040") +
        labs(x = names(dados)[1], y = names(dados)[2])
}
```

Questão 1.

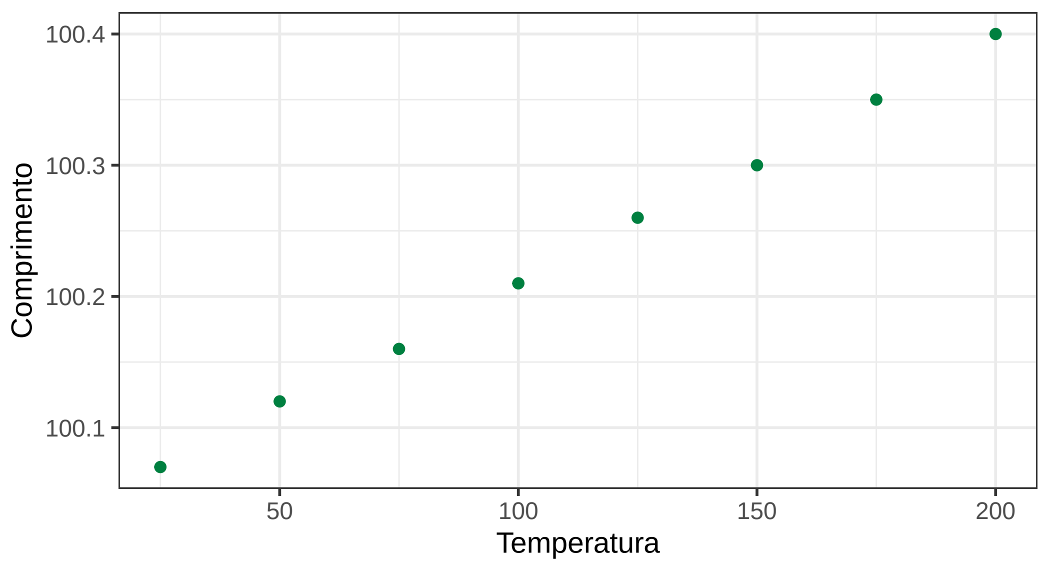
As medições dos comprimentos (y) de uma barra metálica em oito temperaturas (x) diferentes deram origem à tabela abaixo. Para o conjunto de pontos dados:

```
q1 <- tibble(
    Temperatura = c(    25,     50,     75,    100,    125,    150,    175,    200),
    Comprimento = c(100.07, 100.12, 100.16, 100.21, 100.26, 100.30, 100.35, 100.40)
)
```

a) Trace o diagrama de dispersão.
b) Determine a curva de ajuste.
c) Calcule o desvio total.
d) Qual valor estimado do comprimento da barra para 36 °C e 220 °C.

Diagrama de disperção:

```
ggplot(q1, aes(Temperatura, Comprimento)) + geom_point(color = "#008040")
```



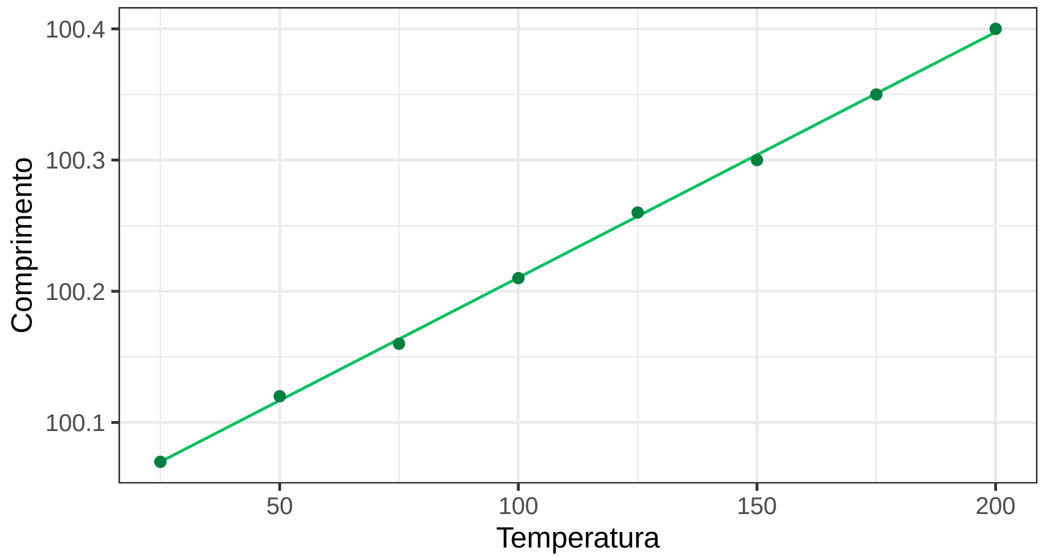
Desvio total e curva de ajuste:

```r
min_quad(q1)
```

alpha	beta	desvio_total	R^2
100.0232	0.0018714	5.36e-05	0.9994176

$$f(x) = 100.023 + 0.002x$$

```r
min_quad_plot(q1)
```



Valor estimado do comprimento da barra para 36 °C e 220 °C:

```r
tibble(
    Temperatura            = c(36, 220),
    `Comprimento estimado` = min_quad_fun(q1)(Temperatura)
)
```

Temperatura	Comprimento estimado
36	100.0906
220	100.4349

Questão 2.

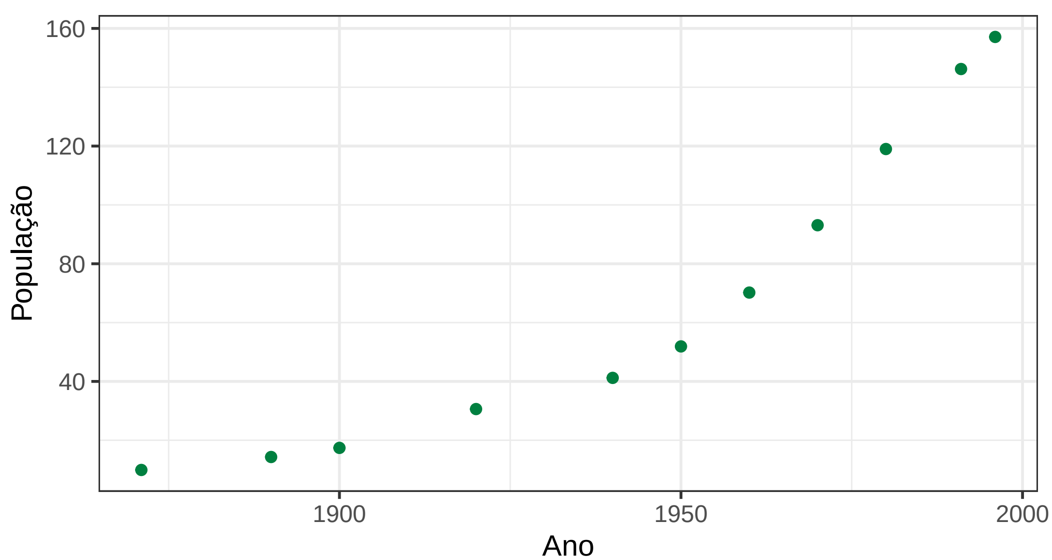
A tabela abaixo apresenta, segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), alguns valores da população brasileira (em milhões de habitantes) (y) e seus respectivos anos (x) de referência.

```
q2 <- tibble(
    Ano       = c(1871, 1890, 1900, 1920, 1940, 1950, 1960, 1970,  1980,  1991,  1996),
    População = c( 9.9, 14.3, 17.4, 30.6, 41.2, 51.9, 70.2, 93.1, 119.0, 146.2, 157.1)
)
```

a) Trace o diagrama de dispersão do conjunto de dados.
b) Ajuste o conjunto de dados a uma função quadrática.
c) Ajuste o conjunto de dados a uma função exponencial.
d) Calcule o desvio nos itens b e c.
e) Estime o valor da população Brasileira nos anos de 2000, 2005 e 2014 segundo os modelos obtidos nos itens b e c e compare-os com os números oficiais, fornecidos pelo IBGE, que são 169.8, 184.2 e 202.7 milhões, respectivamente.

Diagrama de disperção:

```
ggplot(q2, aes(Ano, População)) + geom_point(color = "#008040")
```



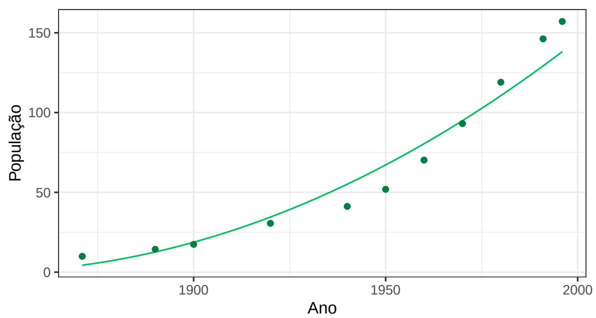
Ajuste quadrático

```r
min_quad(q2, caso = "quadrático") |> round(3)
```

alpha	beta	desvio_total	R²
-142.673	0.077	1301.613	0.956

$$f(x) = -142.673 + 0.077x$$

```r
min_quad_plot(q2, caso = "quadrático")
```



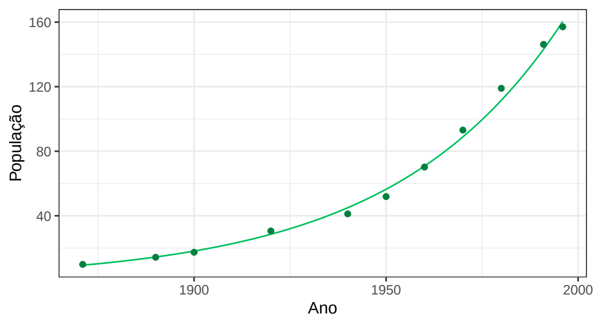
Ajuste exponencial

```r
min_quad(q2, caso = "exponencial") |> round(3)
```

alpha	beta	desvio_total	R²
-40.244	0.023	133.592	0.997

$$f(x) = e^{-40.244 + 0.023x}$$

```r
min_quad_plot(q2, caso = "exponencial")
```



O ajuste exponencial possui menor desvio total: $1301.613 > 133.592$

Valor estimado da população Brasileira nos anos de 2000, 2005 e 2014:

```r
tibble(
    Ano                      = c(2000, 2005, 2014),
    `População estimada`   = min_quad_fun(q2, caso = "exponencial")(Ano),
    `População verdadeira` = c(169.8, 184.2, 202.7)
)
```

Ano	População estimada	População verdadeira
2000	175.6635	169.8
2005	196.7828	184.2
2014	241.4000	202.7

Questão 3.

O banco de dados, ver anexo (Banco de Dados 11.csv) contém informações de 200 CDs comercializados por uma gravadora. Utilize MMQ para verificar se o gasto em publicidade é capaz de prever a venda de CDs. Justifique sua resposta, e também calcule o desvio (D).

```
q3 <- read_csv2("Banco_de_Dados_11.csv")
```

```
min_quad(q3) |> round(3)
```

alpha	beta	desvio_total	R²
125.18	0.105	669141.9	0.363

O gasto em publicidade por si só não permite uma boa previsão da venda de CDs, o R^2, que para um modelo linear simples de duas variáveis é a correlação amostral é de 0.363, que significa que o gasto em publicidade prevê apenas 36.3% da variabilidade da venda de CDs. os dados estão dispersos de mais para usar apenas o gasto em publicidade como variável independente:

```
min_quad_plot(q3)
```