

3º Estágio - Atividade Prática - Estatística Aplicada

Grupo:

- 1 - Lucas de Lima da Silva
- 2 - João Pedro Campos Porto
- 3 - Eduardo Augusto Andrade Bezerra Cavalcanti
- 4 - Erick Farias de Almeida Pereira
- 5 - José Arthur Soares Bezerra

Dataset escolhido: Orange (Dataset padrão do R)

	Tree	age	circumference
1	1	118	30
2	1	484	58
3	1	664	87
4	1	1004	115
5	1	1231	120
6	1	1372	142
7	1	1582	145
8	2	118	33
9	2	484	69
10	2	664	111
11	2	1004	156
12	2	1231	172
13	2	1372	203
14	2	1582	203
15	3	118	30
16	3	484	51
17	3	664	75
18	3	1004	108
19	3	1231	115
20	3	1372	139
21	3	1582	140
22	4	118	32
23	4	484	62
24	4	664	112
25	4	1004	167
26	4	1231	179
27	4	1372	209
28	4	1582	214
29	5	118	30
30	5	484	49
31	5	664	81
32	5	1004	125
33	5	1231	142
34	5	1372	174
35	5	1582	177

1. a) Médias:

Idade: 922,1429

Circunferência: 115,8571

Medianas:

Idade: 1004

Circunferência: 115

Modas:

Idade: 118

Circunferência: 30

Variância:

Idade: 241930,7

Circunferência: 3304,891

Desvio Padrão:

Idade: 491,8645

Circunferência: 57,4882

Coeficiente de correlação: 0,9135

```
Média da circunferência: 115.8571
Média da idade: 922.1429
Variância da circunferência: 3304.891
Variância da idade: 241930.7
Desvio padrão da circunferência: 57.48818
Desvio padrão da idade: 491.8645
Coeficiente de correlação entre idade e circunferência: 0.9135189
Moda da idade: 118
Moda da circunferência: 30
Mediana da idade: 1004
Mediana da circunferência: 115
```

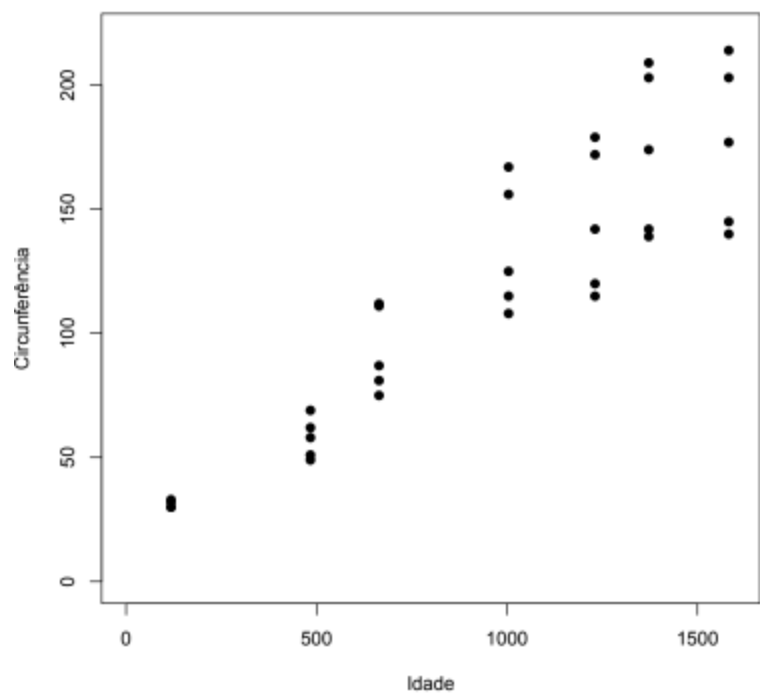
```

# Carregue o dataset Orange (ele já está incluído no R)
data(Orange)
# Calcule a média da coluna "circumference"
media_circumference <- mean(Orange$circumference)
# Calcule a média da coluna "age"
media_age <- mean(Orange$age)
# Calcule a variância da coluna "circumference"
variância_circumference <- var(Orange$circumference)
# Calcule a variância da coluna "age"
variância_age <- var(Orange$age)
# Calcule o desvio padrão da coluna "circumference"
desvioPadrao_circumference <- sd(Orange$circumference)
# Calcule o desvio padrão da coluna "age"
desvioPadrao_age <- sd(Orange$age)
# Calcule o coeficiente de correlação
correlacao <- cor(Orange$age, Orange$circumference)
# Calcule a moda da coluna "age"
tabela_frequenciaAge <- table(Orange$age)
modaAge <- as.numeric(names(tabela_frequenciaAge)[which.max(tabela_frequenciaAge)])
# Calcule a moda da coluna "circumference"
tabela_frequenciaCircumference <- table(Orange$circumference)
modaCircumference <- as.numeric(names(tabela_frequenciaCircumference)[which.max(tabela_frequenciaCircumference)])
# Calcule a mediana da coluna "age"
medianaAge <- median(Orange$age)
# Calcule a mediana da coluna "circumference"
medianaCircumference <- median(Orange$circumference)
# Imprima os valores
cat("Média da circunferência:", media_circumference, "\n")
cat("Média da idade:", media_age, "\n")
cat("Variância da circunferência:", variância_circumference, "\n")
cat("Variância da idade:", variância_age, "\n")
cat("Desvio padrão da circunferência:", desvioPadrao_circumference, "\n")
cat("Desvio padrão da idade:", desvioPadrao_age, "\n")
cat("Coeficiente de correlação entre idade e circunferência:", correlacao, "\n")
cat("Moda da idade:", modaAge, "\n")
cat("Moda da circunferência:", modaCircumference, "\n")
cat("Mediana da idade:", medianaAge, "\n")
cat("Mediana da circunferência:", medianaCircumference, "\n")

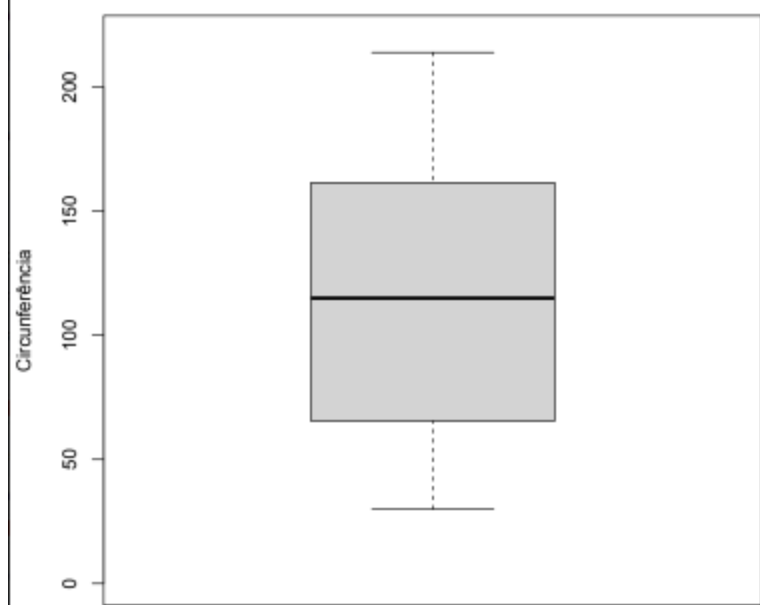
```

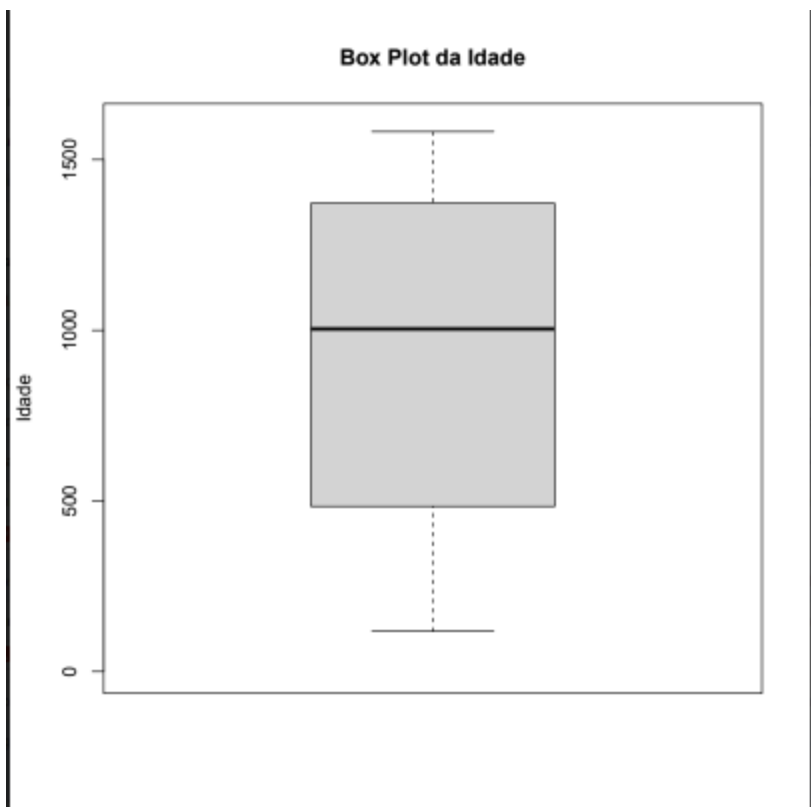
b)

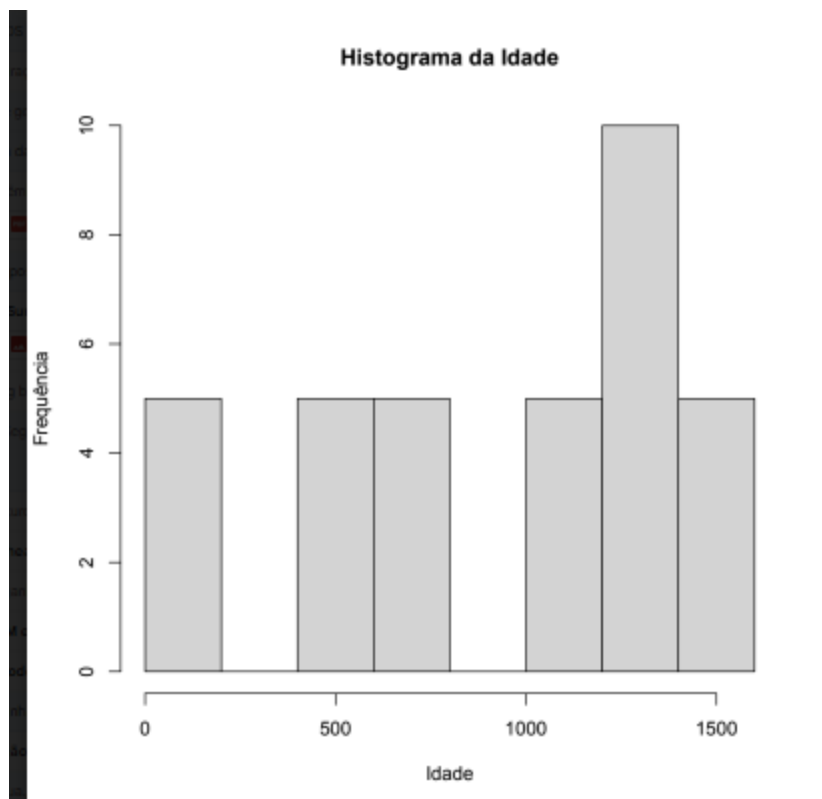
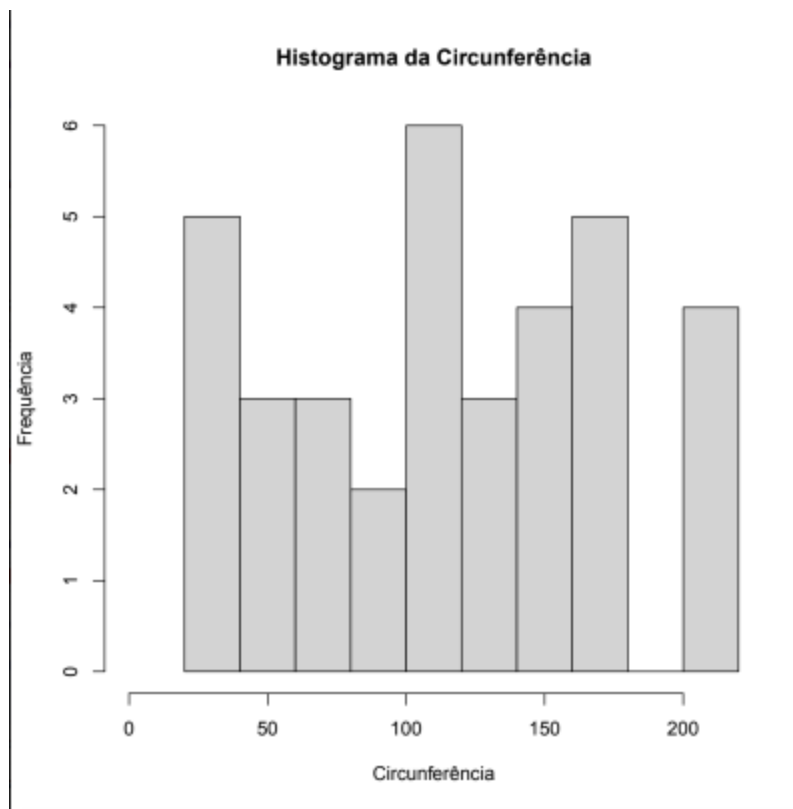
Gráfico de Dispersão Idade vs. Circunferência



Box Plot da Circunferência







2)

a)

```
Intervalo de confiança para a média da Idade (95%): [ 916.9294 , 927.3564 ]
Intervalo de confiança para a média da Circunferência (95%): [ 115.2478 , 116.4664 ]
Intervalo de confiança para a variância (95%): [ 139919.6 , 516103.4 ]
```

```
1 # Função para calcular o intervalo de confiança para a média
2 calcular_intervalo_confianca_media <- function(media_amostra, desvio_padrao, n, nivel_confianca) {
3   z <- qnorm(1 - nivel_confianca / 2) # Valor z correspondente ao nível de confiança
4   erro_padrao <- desvio_padrao / sqrt(n)
5   limite_inferior <- media_amostra - z * erro_padrao
6   limite_superior <- media_amostra + z * erro_padrao
7   return(c(limite_inferior, limite_superior))
8 }
9 # Supondo as estatísticas fornecidas como 95% = 0.95
10 media_idade <- 922.1429
11 desvio_padrao_idade <- 491.8645
12 n_idade <- 35
13 nivel_confianca <- 0.95 # Correção do nível de confiança para 95%
14 media_circunferencia <- 115.8571
15 n_circunferencia <- 35
16 desvio_padrao_circunferencia <- 57.4882
17 # Calcular intervalo de confiança para a média da Idade
18 intervalo_idade <- calcular_intervalo_confianca_media(media_idade, desvio_padrao_idade, n_idade, nivel_confianca)
19 # Calcular intervalo de confiança para a média da Circunferência
20 intervalo_circunferencia <- calcular_intervalo_confianca_media(media_circunferencia, desvio_padrao_circunferencia, n_circunferencia,
21 nivel_confianca)
22 # Exibir os resultados
23 cat("Intervalo de confiança para a média da Idade (95%): [", intervalo_idade[1], ", ", intervalo_idade[2], "]\n")
24 cat("Intervalo de confiança para a média da Circunferência (95%): [", intervalo_circunferencia[1], ", ", intervalo_circunferencia[2], "]\n")
25 intervalo_confianca_variancia <- function(n, variancia_amostra, nivel_confianca) {
26   alpha <- 1 - nivel_confianca
27   graus_liberdade <- n - 1
28   chi_quad_alpha_meio <- qchisq(alpha / 2, df = graus_liberdade)
29   chi_quad_1_menos_alpha_meio <- qchisq(1 - alpha / 2, df = graus_liberdade)
30   limite_inferior <- (n - 1) * variancia_amostra / chi_quad_1_menos_alpha_meio
31   limite_superior <- (n - 1) * variancia_amostra / chi_quad_alpha_meio
32   return(list(limite_inferior = limite_inferior, limite_superior = limite_superior))
33 }
34 n <- 20 # Tamanho da amostra
35 variancia_amostra <- 241930.7 # Variância da amostra
36 nivel_confianca <- 0.95 # Correção do nível de confiança para 95%
37 resultado_intervalo <- intervalo_confianca_variancia(n, variancia_amostra, nivel_confianca)
38 # Exibir os resultados
39 cat("Intervalo de confiança para a variância (95%): [", resultado_intervalo$limite_inferior, ", ", resultado_intervalo$limite_superior, "]\n")
```


3)

a)

Teste de Hipóteses sobre a Média da Resposta (usando o teste t):

Este código realiza um teste estatístico para verificar se a média de uma variável (neste caso, **age** no conjunto de dados **Orange**) é significativamente diferente de um valor específico (100 neste exemplo). O teste t é utilizado

para comparar a média amostral com a média esperada (hipótese nula) e determinar se há evidências estatísticas para afirmar que a média é diferente do valor esperado. O código calcula a estatística de teste t, o valor-p (probabilidade associada à estatística de teste) e conclui se a hipótese nula deve ser rejeitada ou não, com base no nível de significância definido (geralmente 0.05).



```
1 # Carregar o conjunto de dados Orange
2 library(datasets)
3 dados <- Orange
4
5 # Teste de Hipóteses sobre a Média da Resposta
6 teste_media <- t.test(dados$age, mu = 100)
7
8 # Exibir o resultado do teste de hipóteses sobre a média
9 cat("Resultado do Teste de Hipóteses sobre a Média da Resposta
10 (H0: Média = 100):\n")
11 cat("Estatística de Teste:", teste_media$statistic, "\n")
12 cat("Valor-p:", teste_media$p.value, "\n")
13 cat("Conclusão:", ifelse(teste_media$p.value < 0.05, "Rejeitar
14 H0", "Não rejeitar H0"), "\n")
```

Resultado do Teste de Hipóteses sobre a Média da Resposta (H0: Média = 100):
Estatística de Teste: 9.888623
Valor-p: 1.552848e-11
Conclusão: Rejeitar H0

b)

Teste de Hipóteses sobre a Variância da Resposta (usando o teste F manualmente):

Neste código, estamos interessados em verificar se a variabilidade (ou dispersão) de uma variável (novamente, **age** no conjunto de dados **Orange**) é significativamente diferente de um valor específico (50 neste exemplo). Utilizamos um teste F para comparar a variância amostral com a variância esperada (hipótese nula) e determinar se há evidências estatísticas para afirmar que a variância é diferente do valor esperado. O código calcula a estatística de teste F, o valor-p e conclui se a hipótese nula deve ser rejeitada ou não, com base no nível de significância definido (geralmente 0.05).


```
main.r x + ... > Console x Shell x + ...
main.r
1 # Carregar o conjunto de dados Orange
2 library(datasets)
3 dados <- Orange
4
5 # Variância esperada (hipótese nula)
6 variancia_esperada <- 50
7
8 # Número de observações
9 n <- length(dados$age)
10
11 # Variância amostral
12 variancia_amostral <- var(dados$age)
13
14 # Estatística de teste (teste F)
15 estatistica_teste <- ((n - 1) * variancia_amostral) /
    variancia_esperada
16
17 # Graus de liberdade
18 df1 <- n - 1
19 df2 <- Inf
20
21 # Valor p usando a distribuição F
22 valor_p <- 1 - pf(estatistica_teste, df1, df2)
23
24 # Valor p usando a distribuição F
25 valor_p <- 1 - pf(estatistica_teste, df1, df2)
26
27 # Nível de significância
28 nivel_de_significancia <- 0.05
29
30 # Conclusão do teste
31 cat("Resultado do Teste de Hipóteses sobre a Variância da Resposta (H0: Variância = 50):\n")
32 cat("Estatística de Teste:", estatistica_teste, "\n")
33 cat("Valor-p:", valor_p, "\n")
34 cat("Conclusão:", ifelse(valor_p < nivel_de_significancia,
    "Rejeitar H0", "Não rejeitar H0"), "\n")
35
```

623ms on 19:00:26, 11/05 ✓

Resultado do Teste de Hipóteses sobre a Variância da Resposta (H0: Variância = 50):
Estatística de Teste: 164512.9
Valor-p: 0
Conclusão: Rejeitar H0

4. a) Teste de Kolmogorov-Smirnov para 'Age':

Statistic = 0.163547

P-valor = 0.3064295

O valor-p do teste de Kolmogorov-Smirnov (0.3064295) é maior que um nível de significância de 0.05, o que sugere que não há evidências suficientes para rejeitar a hipótese nula de normalidade. Portanto, os dados "Age" podem ser considerados aproximadamente normalmente distribuídos.

Teste de Kolmogorov-Smirnov para 'Circumference':

Statistic = 0.08493812

P-valor = 0.9623232

O valor-p do teste de Kolmogorov-Smirnov (0.9623232) é maior que 0.05. Isso sugere que não há evidências suficientes para rejeitar a hipótese nula de normalidade para a variável "Circumference".

```
1 dados_age = c(Orange$age)
2 dados_circumference = c(Orange$circumference)
3
4 resultado_ks_age <- ks.test(dados_age, "pnorm", mean = mean
  (dados_age), sd = sd(dados_age))
5 resultado_sw_age <- shapiro.test(dados_age)
6
7 resultado_ks_circumference <- ks.test(dados_circumference,
  "pnorm", mean = mean(dados_circumference), sd = sd
  (dados_circumference))
8 resultado_sw_circumference <- shapiro.test(dados_age)
9
10 # Exibir apenas os resultados
11 cat("Teste de Kolmogorov-Smirnov para 'Age':\n", "Statistic =",
  resultado_ks_age$statistic, "\n", "P-valor =",
  resultado_ks_age$p.value, "\n\n")
12 cat("Teste de Shapiro-Wilk para 'Age':\n", "Statistic =",
  resultado_sw_age$statistic, "\n", "P-valor =",
  resultado_sw_age$p.value, "\n\n")
13
14 cat("Teste de Kolmogorov-Smirnov para 'Circumference':\n",
  "Statistic =", resultado_ks_circumference$statistic, "\n", "P-
  -valor =", resultado_ks_circumference$p.value, "\n\n")
15 cat("Teste de Shapiro-Wilk para 'Circumference':\n", "Statistic
  =", resultado_sw_circumference$statistic, "\n", "P-valor =",
  resultado_sw_circumference$p.value, "\n")
```

b) Teste de Correlação de Pearson:

Correlação: 0.9135189

Valor-P: 1.930596e-14

Com base nos resultados do teste de correlação de Pearson:

- A correlação entre as variáveis 'age' e 'circumference' é de aproximadamente 0,9135. Essa é uma correlação positiva forte, o que significa que essas variáveis tendem a aumentar juntas. Em

outras palavras, à medida que a idade das laranjas aumenta, o tamanho da circunferência delas tende a aumentar.

- O valor-p é muito baixo, próximo a 0 ($1.930596e-14$), o que indica que a correlação é estatisticamente significativa. Isso sugere que a correlação observada não é resultado do acaso, mas sim uma relação real entre as variáveis.

```
1 cor_test_result <- cor.test(Orange$age, Orange$circumference,  
  method = "pearson")  
2  
3 cat("Teste de Correlação de Pearson - Resultado:\n")  
4 cat("Correlação:", cor_test_result$estimate, "\n")  
5 cat("Valor-P:", cor_test_result$p.value, "\n")
```

5. a) Estimação pontual:

Intercept: 17.3996502

Idade: 0.1067703

O código usado foi esse:

```
dados <- data.frame(
  idade = c(Orange$age),
  circunferencia = c(Orange$circumference)
)

modelo <- lm(circunferencia ~ idade, data = dados)

coef(modelo)
```

A estimação pontual do intercept ser de 17.3996502 significa que em um caso onde a idade da árvore seja 0 (uma interpretação teórica já que isso dificilmente se aplicaria na vida real) estima-se que sua circunferência será de 17.3996502 mm. Já a estimação pontual de idade ser de 0.1067703 significa que a cada dia que passa é esperado que a circunferência aumente em 0.1067703 mm

Estimação por intervalo:

Intercept: [-0.14328303, 34.9425835]

Idade: [0.08993141, 0.1236092]

Este foi o código usado

```
idade <- c(Orange$age)
circunferencia <- c(Orange$circumference)

dados <- data.frame(idade, circunferencia)

modelo <- lm(circunferencia ~ idade, data = dados)

confint(modelo)
```

O resultado do intercept significa que, com 95% de confiança, acredita-se que o verdadeiro valor do Intercept está dentro desse intervalo. Já o resultado de Idade significa que, com 95% de confiança, acredita-se que a verdadeira mudança média na circunferência para cada unidade adicional de idade está dentro desse intervalo.

b) Variância do erro ε : 546,9042

Utilizei para tal o código em R:

```
data(Orange)
modelo <- lm(circumference ~ age, data = Orange)
residuos <- residuals(modelo)
variancia_erro <- var(residuos)
cat("Variância do erro  $\epsilon$ :", variancia_erro, "\n")
```

c) Coeficiente de determinação (R^2)

No caso do dataset Orange, para obtermos o **coeficiente de determinação** precisamos primeiro obter o coeficiente de correlação entre a idade da árvore (age) e a circunferência da árvore (circumference). Em R isso pode ser feito da seguinte forma:

```
cor(Orange$age, Orange$circumference)
```

Assim obtendo o coeficiente de correlação igual a 0,9135189.

Sabemos que o **coeficiente de determinação** é igual ao coeficiente de correlação ao quadrado, ou seja,

$$0,9135189^2 = 0,8345167 \text{ ou } 83,45\%$$

Isso significa que aproximadamente 83,45% das variações de X (nesse caso idade) pode ser explicado pelas variações de Y (circunferencia).

d)Gráfico da reta ajustada:

Codigo usado:

```
X <- c(Orange$age)
Y <- c(Orange$circumference)

dados <- cbind(X,Y)

cor(X,Y)

plot(X, Y, main = "Idade da árvore x Circunferência")
abline(lm(Y~X))

modelo <- lm(Y~X)
```

coef(modelo)

