Statistics for Engineers
TI0111

Homework III

Nome completo:

Vinicius Alexandre Gomes do Nascimento e Matheus Parente Reis

Numero de matricula: 568594 e 571954

Questão 1

Assume-se que o tempo de vida X (medido em anos) de um computador segue uma distribuição exponencial com parâmetro desconhecido $\lambda > 0$. Uma amostra aleatória dos tempos de vida dos computadores é apresentada na Tabela 1. Os dados são fictícios e são utilizados apenas para fins ilustrativos.

0.99	2.31	10.85	6.15	10.81	3.72	5.75	4.15	9.27	7.84
2.31	10.85	6.15	1.81	3.72	5.75	10.40	10.04	4.15	9.27

Tabela 1: Dados usados na questão 1: Tempo de vida (em anos) dos computadores.

1. Escreva a função densidade de probabilidade da distribuição exponencial com parâmetro λ.
2. Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$:
   (a) Escreva a função de verossimilhança $L(\lambda)$.
   (b) Derive a correspondente função log-verossimilhança $\ell(\lambda)$.
   (c) Determine o estimador de máxima verossimilhança (MLE, do inglês) $\hat{\lambda}$ de λ.
3. Utilizando os dados fornecidos na Tabela 1, calcule o valor numérico do MLE $\hat{\lambda}$.
4. Construa o gráfico da função log-verossimilhança $\ell(\lambda)$ com base nos dados observados, considerando um intervalo adequado de valores para λ. Indique claramente no gráfico o valor do estimador de máxima verossimilhança $\hat{\lambda}$.
5. Utilizando o parâmetro estimado $\hat{\lambda}$:
   (a) Calcule o tempo médio de vida estimado de um computador.
   (b) Calcule a probabilidade de que um computador funcione por mais de 5 anos.

6. A distribuição exponencial possui a *propriedade da falta de memória*, o que significa que a probabilidade de falha no futuro não depende do tempo que o computador já esteve em funcionamento.

   (a) Explique essa propriedade com suas próprias palavras.

   (b) Discuta brevemente se essa suposição parece razoável para modelar o tempo de vida de computadores.

Solução da questão 1

Item 1

1.1 Fundamentação

A **distribuição exponencial** possui uma forte relação com a distribuição de poisson. Enquanto com poisson, modela-se a contagem do número de eventos dentro de um intervalo fixo, a exponencial descreve o tempo entre a ocorrência desses eventos.

1.2 Resposta

A função densidade de probabilidade de uma distribuição exponencial com parâmetro λ é dada por:

$$f_X(x) = \lambda e^{-\lambda x}, \text{ onde } 0 \leq x < \infty$$

Item 2

2.1 Fundamentação

A **função de verossimilhança** $L(\theta)$ quantifica a probabilidade de observar uma determinada amostra para diferentes valores do parâmetro θ. Diferente da função de densidade usual, que considera os parâmetros fixos e os dados como variáveis, a verossimilhança trata os dados observados como fixos e a função varia conforme o parâmetro desconhecido.

Para uma amostra de observações independentes e identicamente distribuídas (i.i.d.), a verossimilhança é o produto das densidades individuais:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

O objetivo do método é encontrar o valor de θ que maximiza esta função. Esse valor, que torna a amostra observada a mais provável sob o modelo assumido, é denominado **Estimador de Máxima Verossimilhança** (MLE, do inglês *Maximum Likelihood Estimator*), denotado por $\hat{\theta}$.

Na prática, é mais conveniente trabalhar com a **função de log-verossimilhança** $\ell(\theta)$, definida como:

$$\ell(\theta) = \ln\left(L(\theta)\right)$$

Como o logaritmo é uma função monotonicamente crescente, o valor de θ que maximiza $\ell(\theta)$ é o mesmo que maximiza $L(\theta)$. A vantagem dessa transformação reside na propriedade de converter o produto de densidades em uma soma, simplificando significativamente o cálculo da derivada. O MLE é obtido derivando-se $\ell(\theta)$ em relação a θ e igualando o resultado a zero (condição de primeira ordem).

2.2 Resposta

(a) Com base na fundamentação, tem-se que a função de verossimilhança $L(\lambda)$ é dada por:

$$L(\lambda) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdot \ldots \cdot \lambda e^{-\lambda x_n}$$

$$L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

(b) Sendo aplicada a função $ln()$ em ambos lados se tem a função de log-verossimilhança:

$$\ln L(\lambda) = \ln\left(\lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}\right)$$

$$\ell(\lambda) = \ln\left(\lambda^n\right) \cdot \ln e^{-\lambda \sum_{i=1}^{n} x_i} = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

(c) Com objetivo de encontrar o MLE, primeiro se deriva a função $\ell(\lambda)$ em relação a λ:

$$\ell(\lambda)' = \frac{d(\ell(\lambda))}{d\lambda} = \frac{d(n \ln(\lambda))}{d\lambda} - \frac{d(\lambda \sum_{i=1}^{n} x_i)}{d\lambda}$$

$$\ell(\lambda)' = n \cdot \frac{d(\ln(\lambda))}{d\lambda} - \sum_{i=1}^{n} x_i \frac{d(\lambda)}{d\lambda} = n \cdot \frac{1}{\lambda} - \sum_{i=1}^{n} x_i \cdot 1$$

$$\ell(\lambda)' = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

Restando encontrar o valor de λ tal que $\ell(\lambda)' = 0$

$$\ell(\lambda)' = \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i = 0$$

$$\frac{n}{\hat{\lambda}} = \sum_{i=1}^{n} x_i$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

Considerando que a média amostral é dada por: $\overline{X} = \frac{\sum_{i=1}^{n} x}{n}$. Tem-se que o **estimador de máxima verossimilhança** da distribuição exponencial é o inverso da média aritmética das amostras.

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

Item 3

3.1 Fundamentação

Com base no resultado obtido no item anterior, pode-se encontrar o valor do **estimador de máxima verossimilhança** (MLE) $\hat{\lambda}$ com a expressão:

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

3.2 Resposta

Considerando os 20 valores dados, obtêm-se o valor da média aritmética da amostra:

$$\overline{X} = \frac{0.99 + 2.31 + 10.85 + ... + 9.27}{20} \approx 6.3145$$

Logo,

$$\hat{\lambda} = \frac{1}{\overline{X}} = \frac{1}{6.31545} = 0.15836$$

3.3 Código em R

Código em R que implementa o cálculo do MSE com base na equação $\hat{\lambda} = \frac{1}{\overline{X}}$ e a amostra.

Listado 1: Cálculo do MLE para a Questão 1 Item 3

```
sample = c(0.99, 2.31, 10.85, 6.15, 10.81, 3.72, 5.75, 4.15, 9.27, 7.84,
    2.31, 10.85, 6.15, 1.81, 3.72, 5.75, 10.40, 10.04, 4.15, 9.27) #
    Amostra

m = mean(sample) # Calculo da media

mle = 1/m

mle

# Saida
[1] 0.1583657
```

ITEM 4

4.1 Gráfico

Utilizando R para geração do gráfico (código mais abaixo), com base na amostra.

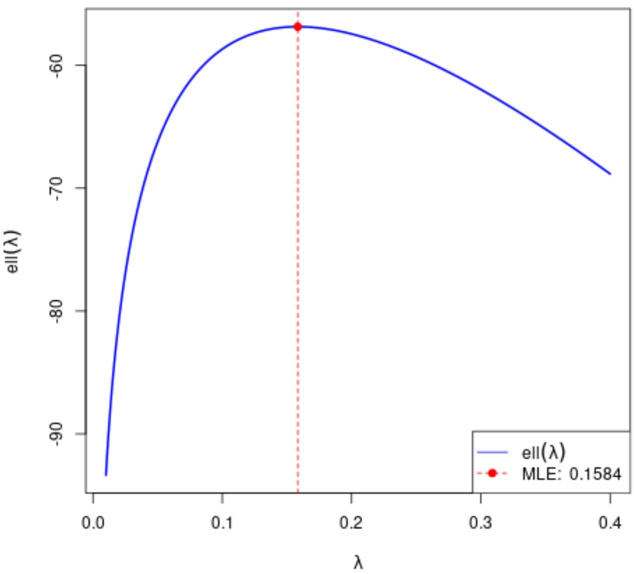

Figura 1: Gráfico Questão 1 Item 4

Em relação ao gráfico (Figura 1), o intervalo utilizado foi pensado para englobar o valor de MLE sem estender desnecessariamente o gráfico. Sendo possível observar de forma clara o comportamento da função de **log-verossimilhança** $\ell(\lambda)$ e que o valor de MLE $\hat{\lambda} = 0.158$ proposto corresponde ao seu máximo.

4.2 Código em R

Código utilizado para gerar o gráfico.

Listado 2: Gerar Grafico Questao 1 Item 4

```
# Dados da amostra
dados <- c(0.99, 2.31, 10.85, 6.15, 10.81, 3.72, 5.75, 4.15, 9.27, 7.84,
           2.31, 10.85, 6.15, 1.81, 3.72, 5.75, 10.40, 10.04, 4.15, 9.27)

# Definir parametros basicos
n <- length(dados) # Tamanho da amostra
soma_x <- sum(dados) # Total da soma
lambda_hat <- n / soma_x  # Valor do MLE calculado no item 3

# Definir a funcao de Log-Verossimilhanca
log_verossimilhanca <- function(l) {
  return(n * log(l) - l * soma_x)
}

# Intervalo de valores para lambda ao redor do MLE para o grafico
vetor_lambda <- seq(0.01, 0.4, length.out = 1000)
valores_log_v <- log_verossimilhanca(vetor_lambda)

# Construir o grafico
plot(vetor_lambda, valores_log_v, type = "l", col = "blue", lwd = 2,
     xlab = expression(lambda), ylab = expression(ell(lambda)))

# Indicar o valor do MLE no grafico
abline(v = lambda_hat, col = "red", lty = 2) # Linha vertical no MLE
points(lambda_hat, log_verossimilhanca(lambda_hat), col = "red", pch =
    19) # Ponto no topo

# Adicionar legenda
legend("bottomright", legend = c(expression(ell(lambda)), paste("MLE:␣",
    round(lambda_hat, 4))),
       col = c("blue", "red"), lty = c(1, 2), pch = c(NA, 19))
```

Item 5

5.1 Fundamentação

(a) O valor do **tempo médio** $E(X)$ para a distribuição exponencial é dada por:

$$E(X) = \frac{1}{\hat{\lambda}}$$

Observando a fundamentação do **Item 3**, onde $\hat{\lambda} = \frac{1}{\overline{X}}$ nota-se que o resultado para o valor médio da distribuição exponencial é exatamente o valor médio das amostras $E(X) = \overline{X}$.

(b) Para calcular a probabilidade de P(X > x) utiliza-se o complemento da CDF da exponencial $F(x) = 1 - e^{-\lambda x}$:

$$P(X > x) = 1 - P(X \leq x) = 1 - F(x) = 1 - 1 + e^{-\lambda x}$$

$$P(X > x) = e^{-\lambda x}$$

5.2 Resposta

(a) Utilizando da primeira parte da fundamentação se tem que o tempo médio da exponencial com o parâmetro lambda $\hat{\lambda}$ =:

$$E(X) = \frac{1}{\hat{\lambda}} = \overline{X} = 6.3145$$

Logo, o tempo médio de vida para um computador é de 6.3145 anos, ou cerca de 6 anos e 114 dias.

(b) Com base na equação apresentada na fundamentação tem se que a probabilidade de um computador funcionar por mais de 5 anos $P(X > 5)$ é dada por:

$$P(X > 5) = e^{-\hat{\lambda} \cdot 5} = e^{-0.15836 \cdot 5} = e^{-0.7918} = 0.4530 = 45.30\%$$

Dessa forma, a probabilidade para um computador ter seu tempo de vida maior que 5 anos é cerca de 45.30%.

5.3 Código em R

Código em R que faz o calculo dos itens (a) e (b)

Listado 3: Questão 1 item 5

```
# Dados da amostra
dados <- c(0.99, 2.31, 10.85, 6.15, 10.81, 3.72, 5.75, 4.15, 9.27, 7.84,
           2.31, 10.85, 6.15, 1.81, 3.72, 5.75, 10.40, 10.04, 4.15, 9.27)

# Calculo do Lambda
# O MLE de lambda para a exponencial e 1 / media
lambda_mle <- 1 / mean(dados)


# Item (a) Tempo medio de vida estimado
tempo_medio <- mean(dados) # Assim como definido na fundamentacao

tempo_medio # Apresenta o valor

# Item (b) P(X > 5) com a funcao de distribuiccao acumulada pexp
# lower.tail = FALSE calcula diretamente a probabilidade de ser MAIOR que
    5
prob <- pexp(5, rate = lambda_mle, lower.tail = FALSE)

prob # Apresenta o valor

# Saida
[1] 6.3145 # Valor do tempo medio E(X)
[1] 0.4530158 # Valor de P(X>5) probabilidade do tempo de vida ser acima
    de 5 anos
```

ITEM 6

(a) A propriedade tem sentido de que o o cálculo da probabilidade de falha nos próximos anos não considera o tempo que já foi passado anteriormente, ou seja, a probabilidade de falha dentro de t anos para um computador novo e um que já esta funcionando a s anos é a mesma considerando a distribuição exponencial, não representando o desgaste do computador, de certa forma como se "sempre estivesse novo".

(b) Nesse sentido, a distribuição exponencial não parece ser suficiente para modelar o tempo de vida de computadores, pois ao passar do tempo existe uma degradação natural dos componentes do computador que fazem com que computadores mais antigos sejam mais propensos a falhas, aumentando sua chance de falha no futuro quando comparado com outros computadores mais novos.

Questão 2

O conjunto de dados de `penguins`, na biblioteca `palmerpenguins`[1] do R, contém medidas para as três espécies de pinguins (figura 3): ilha no arquipélago Palmer na Antártica, tamanho (comprimento da nadadeira, massa corporal, dimensões do bico) e sexo. Importe o conjunto de dados[2] e familiarize com ele.

Figura 2: Espécies e características dos pinguins na questão 2.

1. Considere a massa corporal (`body_mass`) em gramas como variável independente, x, e o comprimento do bico (`bill_length`) em milimetras como variável dependente y. Construa um gráfico de dispersão entre x and y. Com base no gráfico, comente se uma relação linear entre as variáveis parece plausível.

2. Defina os parâmetros da reta de regressão com o método dos mínimos quadrados e verifique os resultados obtidos com o comando `lm()` no R. Adicione a reta de regressão no gráfico de dispersão.

3. Calcule os resíduos da regressão e apresente uma representação gráfica dos mesmos. Em seguida, calcule a raiz do erro quadrático médio (RMSE, do inglês) e o coeficiente de determinação R^2. Comente sobre os resultados obtidos.

4. O conjunto de dados não apresenta outliers evidentes. Modifique esse conjunto introduzindo artificialmente uma observação extrema, seja por meio de um aumento ou de uma redução substancial no valor da massa corporal ou do comprimento do bico de um dos pinguins. Em seguida, ajuste um modelo de regressão linear utilizando o conjunto de dados modificado. Compare os coeficientes estimados da regressão, as retas ajustadas e os valores do RMSE e do R^2 com aqueles obtidos no item 2. Por fim, discuta a influência da observação artificialmente introduzida sobre os resultados da regressão.

---

[1] https://cran.r-project.org/web/packages/palmerpenguins/index.html

[2] `install.packages("palmerpenguins"); library(palmerpenguins);`
`penguins_data <-na.omit(penguins) # desconsiderando os dados faltantes`

Solução da questão 2

A tarefa propõe a análise de dados (amostras) coletados de pinguins que habitam o Arquipélago de Palmer, na Antártica. Essa análise consiste em traçar um paralelo (correlação linear) entre o comprimento do bico da ave e sua massa corporal.

Uma das possíveis motivações dessa correlação é a percepção dos biólogos de uma proporção direta do comprimento do bico em relação à massa corporal. Isso decorre, possivelmente, do fato de aves com bicos maiores poderem se alimentar de forma mais eficiente, ou qualquer outra hipótese que venha a ser provada estatisticamente.

Item 1

Primeiramente, para construção e validação da hipótese, utiliza-se o seguinte script em R:

Listado 4: Dispersão dos dados de Tamanho do bico por Massa corporal

```
# Importacao das amostras
library(palmerpenguins)

# Remover linhas com valores ausentes
penguins <- na.omit(penguins)

# Definir x e y para facilitar o uso nas formulas
x <- penguins$body_mass_g
y <- penguins$bill_length_mm

# Plot do grafico de Dispersao das amostras
plot(x, y,
     main = "Massa Corporal vs Comprimento do Bico",
     xlab = "Massa Corporal (g)",
     ylab = "Comprimento do Bico (mm)",
     pch = 19,
     col = "darkblue")
```

Gerando o seguinte gráfico:

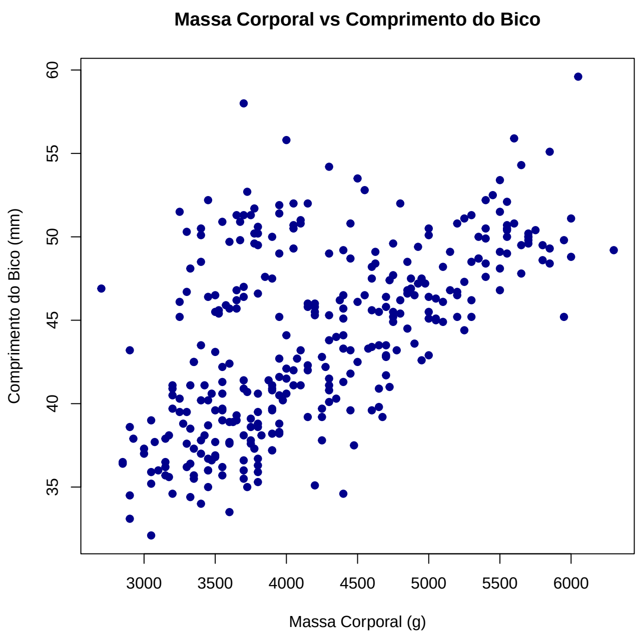

Figura 3: Espécies e características dos pinguins na questão 2.

Os dados, apesar de dispersos, demonstram um certo aumento no comprimento do bico conforme a massa corporal aumenta, logo, apesar de imperfeita, denota-se uma correlação linear positiva entre esses dois dados.

Essa correlação também pode ser obtida por meio da fórmula matemática:

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

Que na linguagem R é calculada da seguinte forma:

Listado 5: Cálculo do coeficiente de correlação

```
# Calculando a correlacao de Pearson entre massa e comprimento do bico
correlacao <- cor(x, y)
```

Resultando em, aproximadamente, $r = 0.5895$. Esse dado é considerado uma **relação moderada**, o que já descreve uma certa margem de erro do modelo linear em relação a dispersão original.

Item 2

A equação da reta estimada é definida por $\hat{y} = \beta_0 + \beta_1 x$, onde a notação $\hat{y}$ indica que o valor gerado pela equação linear é uma predição. Os parâmetros β_0 e β_1 são selecionados de tal forma que a soma dos quadrados dos desvios seja minimizada, garantindo o melhor ajuste linear aos dados.

Segue uma breve apresentação intuitiva dos parâmetros:

- O intercepto (β_0) representa o valor base teórico do comprimento do bico quando a massa corporal é nula;
- O coeficiente angular (β_1) indica a taxa de crescimento proporcional, ou seja, o quanto o bico aumenta para cada unidade de massa corporal acrescida.

Matematicamente, os estimadores de mínimos quadrados são calculados pelas fórmulas:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \tag{2}$$

Em R, a definição dos parâmetros de forma "manual"e a respectiva verificação com o comando `lm()` (*linear model*) são realizadas conforme o bloco abaixo:

Listado 6: Definição de parâmetros e verificação com lm()

```
# Definicao das variaveis
x <- penguins$body_mass_g
y <- penguins$bill_length_mm

# Calculo manual dos parametros
beta1_manual <- sum((x - mean(x)) * (y - mean(y))) / sum((x - mean(x))^2)
beta0_manual <- mean(y) - beta1_manual * mean(x)

# Verificacao com o comando lm()
modelo <- lm(bill_length_mm ~ body_mass_g, data = penguins)
coef(modelo) # Comparacao com os valores manuais
```

Como resultado, obteve-se $\beta_0 \approx 27,2$ mm e $\beta_1 \approx 0,004$ mm/g. A convergência entre o cálculo algébrico e a função nativa do R justifica a utilização do comando `lm()` para as etapas subsequentes. Para a representação gráfica, a reta de regressão foi sobreposta ao gráfico de dispersão:

Listado 7: Adição da reta de regressão ao gráfico

```
plot(penguins$body_mass_g, penguins$bill_length_mm,
     main = "Massa Corporal vs Comprimento do Bico",
     xlab = "Massa Corporal (g)",
     ylab = "Comprimento do Bico (mm)",
     pch = 19, col = "darkblue")

# Adiciona a reta de regressao ao grafico de dispersao
abline(modelo, col = "red", lwd = 2)
```

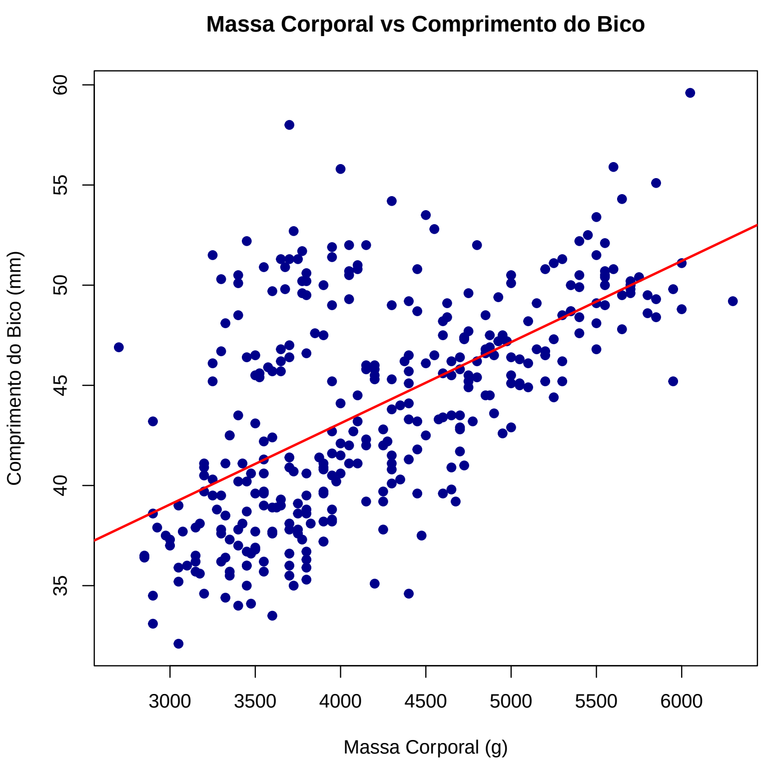

Figura 4: Gráfico de dispersão com a reta de regressão linear sobreposta.

O gráfico evidencia o comportamento linear dos dados, com a reta de regressão representando a tendência central do crescimento do bico conforme o aumento da massa corporal.

Item 3

Os resíduos consistem na diferença individual entre o valor real observado e o valor predito pela reta de regressão (em outras palavras, o erro da predição), evidenciando a defasagem do modelo predito em relação ao original.

Afim de quantificar a eficácia do ajuste, utiliza-se o coeficiente de determinação (R^2), que representa a proporção da variância total dos dados que é explicada pelo modelo linear. Complementarmente, a raiz do erro quadrático médio, ou RMSE, fornece a magnitude média do erro nas mesmas unidades da variável dependente (milímetros).

Matematicamente, estas métricas são dadas pelas seguintes expressões:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

Em R, a implementação para a obtenção dos resultados e do gráfico é realizada conforme o bloco abaixo:

Listado 8: Cálculo de resíduos e métricas de desempenho

```r
# Obtendo os residuos do modelo ajustado
residuos <- residuals(modelo)

# Gerando a representacao grafica dos residuos
plot(predict(modelo), residuos,
     main = "Grafico de Residuos",
     xlab = "Valores Preditos (mm)",
     ylab = "Residuos (mm)",
     pch = 19, col = "darkgreen")
abline(h = 0, col = "red", lty = 2)

# Calculo do RMSE e R-quadrado
rmse <- sqrt(mean(residuos^2))
r_quadrado <- summary(modelo)$r.squared
```

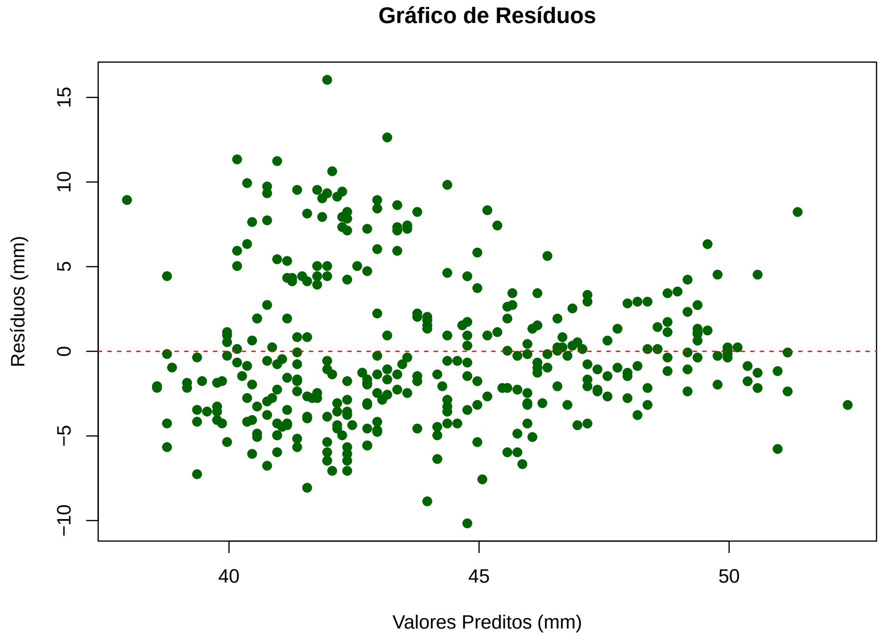

Figura 5: Distribuição dos resíduos em relação aos valores preditos pelo modelo.

O modelo de regressão linear retorna um coeficiente de determinação (R^2) de aproximadamente $0,3475$. Este valor indica que cerca de $34,8\%$ da variabilidade observada no comprimento do bico é explicada pela massa corporal dos pinguins analisados. Ademais, o valor de $4,41$ mm obtido para o RMSE descreve a dispersão média dos dados em relação ao modelo, um resultado condizente com o coeficiente de Pearson moderado calculado anteriormente.

Observando o gráfico de resíduos, nota-se que, apesar do baixo poder de predição indicado pelo R^2, o modelo linear mostra-se minimamente satisfatório. Isso ocorre porque os resíduos oscilam de maneira aleatória em torno de zero, não apresentando padrões claros. Essa aleatoriedade sugere a ausência de tendências não lineares óbvias, indicando que modelos mais complexos não necessariamente ofereceriam uma descrição mais refinada em comparado ao modelo linear.

Item 4

Para este experimento, modificou-se a massa corporal de uma das observações para 10.000 g, um valor significativamente superior, podendo ser considera um outlier

Com a fabricação dessa amostra, obteve-se as seguintes alterações

Os dados da tabela evidenciam que a introdução de uma única observação atípica foi suficiente para provocar uma alteração drástica no comportamento do modelo, resultando em variações significativas tanto na inclinação (β_1) quanto no deslocamento vertical (β_0)

Parâmetro	Modelo Original	Modelo com Outlier	Variação %
Intercepto (β_0)	$\approx 27,22$	$\approx 37,15$	$+36,48$
Inclinação (β_1)	$\approx 0,0040$	$\approx 0,0018$	$-55,00$
$RMSE$	$\approx 4,41$ mm	$\approx 5,20$ mm	$+17,91$
R^2	$\approx 0,3475$	$\approx 0,0512$	$-85,26$

Tabela 2: Comparação dos resultados e variação percentual após a introdução do *outlier*.

da reta de regressão. Este fenômeno decorre diretamente da natureza do método de mínimos quadrados ordinários, que busca minimizar a soma dos resíduos ao quadrado. Como o cálculo do RMSE penaliza severamente desvios de grande magnitude, o modelo torna-se circunstancial ao *outlier*, deslocando a reta na tentativa de aproximá-la desse valor extremo para reduzir o erro global.

Portanto, observa-se uma perda substancial na capacidade preditiva do modelo, percebida pela queda acentuada no coeficiente de determinação (R^2). A eficácia preditiva é comprometida pois a reta deixa de representar a tendência central da massa principal de dados para tentar acomodar o ponto discrepante. Tais resultados são reflexo da alta sensibilidade da regressão linear a valores atípicos e reforçam a importância da aplicação de estratégias de tratamento de dados, como o método do Intervalo Interquartil (IQR). Tais táticas são fundamentais para a identificação e remoção de *outliers*, permitindo a construção de um modelo estatisticamente robusto e com poder preditivo satisfatório.

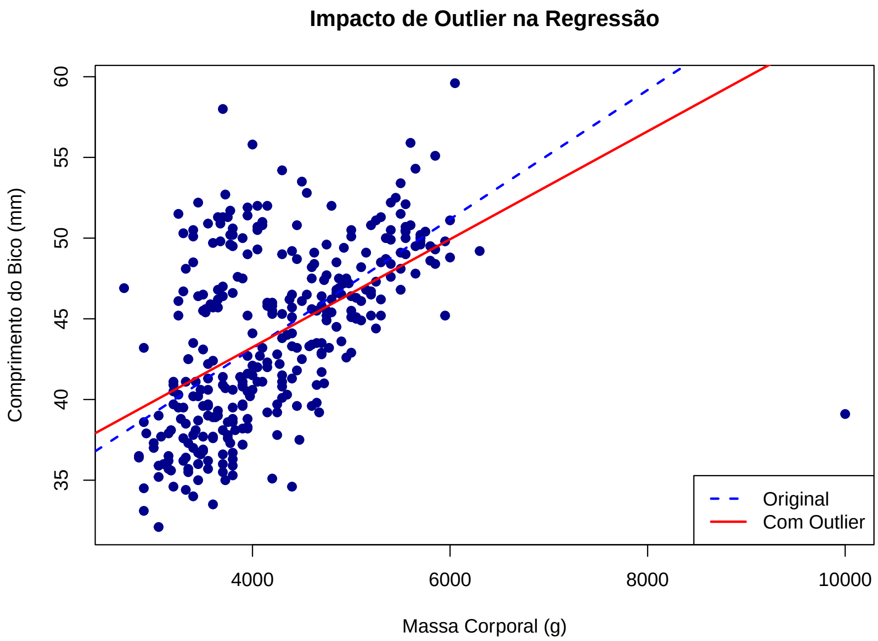

Figura 6: Visualização das retas de regressão original e modificada.