Statistics for Engineers
TI0111

Homework III

Nome completo: Savlio Carvalho Pontes

Número de matricula: 567715

Nome completo: Thaís Sousa Barros Leal

Número de matricula: 565548

Base Teórica

Esta seção apresenta os fundamentos teóricos necessários para o desenvolvimento das análises estatísticas propostas neste trabalho. São abordados conceitos de variáveis aleatórias, distribuições de probabilidade, estimação de parâmetros pelo método da máxima verossimilhança, regressão linear simples e medidas de ajuste de modelos estatísticos. Esses tópicos fornecem a base conceitual para a modelagem do tempo de vida de computadores e para a análise da relação entre características morfológicas de pinguins.

Variáveis Aleatórias

Uma variável aleatória é uma função que associa um valor numérico a cada resultado possível de um experimento aleatório. As variáveis aleatórias podem ser classificadas em dois tipos principais:

- **Variáveis aleatórias discretas**, que assumem valores contáveis, como o número de falhas observadas em um sistema ou o número de indivíduos em uma amostra.
- **Variáveis aleatórias contínuas**, que assumem valores em intervalos reais, como o tempo de vida de um equipamento ou medidas físicas, como comprimento e massa.

Distribuição Exponencial

A distribuição exponencial é amplamente utilizada para modelar o tempo até a ocorrência de um evento, como falhas em equipamentos eletrônicos. Ela é caracterizada por um único parâmetro $\lambda > 0$, denominado taxa de falha. Sua função densidade de probabilidade é dada por

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Uma propriedade fundamental dessa distribuição é a *falta de memória*, que indica que a probabilidade de falha futura é independente do tempo já decorrido de funcionamento do sistema.

Máxima Verossimilhança

O método da máxima verossimilhança é uma técnica de estimação de parâmetros baseada na maximização da probabilidade de observar os dados amostrais. Dada uma amostra aleatória $X_1, X_2, \ldots, X_n$ com função densidade dependente de um parâmetro λ, define-se a função de verossimilhança como

$$L(\lambda) = \prod_{i=1}^{n} f(X_i; \lambda).$$

Em geral, trabalha-se com a função log-verossimilhança, obtida aplicando o logaritmo natural à verossimilhança, pois isso simplifica os cálculos. O estimador de máxima verossimilhança (MLE) é o valor do parâmetro que maximiza essa função.

Regressão Linear Simples

A regressão linear simples é utilizada para modelar a relação entre uma variável explicativa x e uma variável resposta y. O modelo assume a forma

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

onde β_0 é o intercepto, β_1 é o coeficiente angular e ε representa o erro aleatório, assumido com média zero e variância constante. Os parâmetros do modelo são estimados pelo método dos mínimos quadrados, que minimiza a soma dos quadrados dos resíduos.

Medidas de Qualidade do Ajuste

A qualidade do ajuste de um modelo de regressão pode ser avaliada por diferentes métricas. O erro quadrático médio (RMSE) mede a magnitude média dos erros de predição, enquanto o coeficiente de determinação R^2 indica a proporção da variabilidade da variável resposta explicada pelo modelo. Valores elevados de R^2 e baixos de RMSE indicam melhor ajuste aos dados observados.

1 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.

1. Escreva a função densidade de probabilidade da distribuição exponencial com parâmetro λ.

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.

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

1.1) Função densidade de probabilidade.

Assume-se que o tempo de vida X de um computador segue uma distribuição exponencial com parâmetro $\lambda > 0$. A função densidade de probabilidade (PDF) dessa distribuição é dada por

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Essa função descreve a probabilidade associada ao tempo até a falha do computador, assumindo uma taxa de falha constante ao longo do tempo.

```r
    # ----------- QUESTÃO 1 ----------- #
# Considera-se que o tempo de vida segue uma distribuição exponencial com taxa λ > 0

# 1.1) PDF da distribuição exponencial
exp_pdf <- function(x, lambda){
  return(lambda * exp(-lambda * x))
}
```

Listado 1: Função densidade de probabilidade da distribuição exponencial

1.2) Função de verossimilhança, log-verossimilhança e estimador de máxima verossimilhança.

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória independente com distribuição exponencial de parâmetro λ.

(a) Função de verossimilhança

A função de verossimilhança é definida como o produtório das densidades avaliadas nos dados observados:

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

```r
 # 1.2(a) Função de verossimilhança(likelihood)
likelihood_exp <- function(x, lambda) {
  return(prod(exp_pdf(x, lambda)))
}
```

Listado 2: Função de verossimilhança

(b) Função log-verossimilhança

Aplicando o logaritmo natural à função de verossimilhança, obtém-se a função log-verossimilhança:

$$\ell(\lambda) = \log L(\lambda) = n \log(\lambda) - \lambda \sum_{i=1}^{n} X_i.$$

```r
# 1.2(b) Função log-verossimilhança(log likelihood)
loglike_exp <- function(x, lambda) {
  n <- length(x)
  return(n * log(lambda) - lambda * sum(x))
}
```

Listado 3: Função log-verossimilhança

(c) Estimador de máxima verossimilhança

Para determinar o estimador de máxima verossimilhança, deriva-se a função log-verossimilhança em relação a λ e iguala-se a zero:

$$\frac{d\ell(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0.$$

Resolvendo para λ, obtém-se:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\bar{X}},$$

onde $\bar{X}$ representa a média amostral.

```r
 # 1.2(c) Estimador de Máxima Verossimilhança
# Para a exponencial, o MLE é o inverso da média amostral
mle_exp <- function(x){
  return(1 / mean(x))
}
```

Listado 4: Estimador de máxima verossimilhança

1.3) Cálculo do estimador de máxima verossimilhança.

Utilizando os dados apresentados na Tabela 1, calcula-se inicialmente a média amostral $\bar{X}$, dada por

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

onde $n = 20$ é o tamanho da amostra.

Somando os valores observados, obtém-se:

$$\sum_{i=1}^{20} X_i = 126.32.$$

Assim, a média amostral é:

$$\bar{X} = \frac{126.32}{20} = 6.316.$$

O estimador de máxima verossimilhança para o parâmetro λ da distribuição exponencial é dado por:

$$\hat{\lambda} = \frac{1}{\bar{X}}.$$

Substituindo o valor da média amostral, tem-se:

$$\hat{\lambda} = \frac{1}{6.316} \approx 0.15832. \text{(No R: 0.15836)}$$

```r
# 1.3) Dados da amostra
d <- 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
)

# Cálculo do estimador de máxima verossimilhança
estimador <- mle_exp(d)
```

Listado 5: Cálculo do estimador de máxima verossimilhança.

1.4) Gráfico da função log-verossimilhança.

A função log-verossimilhança foi avaliada para um intervalo adequado de valores positivos de λ. O gráfico resultante permite identificar o ponto em que $\ell(\lambda)$ atinge seu valor máximo, correspondente ao estimador de máxima verossimilhança $\hat{\lambda}$. No gráfico, esse valor é indicado por uma linha vertical.

- **Código em R** O código utilizado para gerar o gráfico da função log-verossimilhança encontra-se apresentado no Listado correspondente.

```r
# 1.4) Gráfico da função log-verossimilhança

# Valores de lambda analisados
lambda_vals <- seq(0.01, 1, length.out = 300)

# Cálculo da log-verossimilhança para cada valor de λ
loglike_vals <- sapply(lambda_vals, loglike_exp, x = d)

# Gráfico da log-verossimilhança em função de λ
plot(lambda_vals, loglike_vals, type = "l",
     xlab = expression(lambda),
     ylab = expression(ell(lambda)))

# Linha vertical indicando o valor do MLE
abline(v = estimador, col = "red", lwd = 2, lty = 2)

# Legenda do gráfico
legend("topright",
       legend = expression(hat(lambda)),
       col = "red",
```

```
21          lty = 2,
22          lwd = 2)
```

Listado 6: Gráfico da função log-verossimilhança

- **Gráfico obtido**

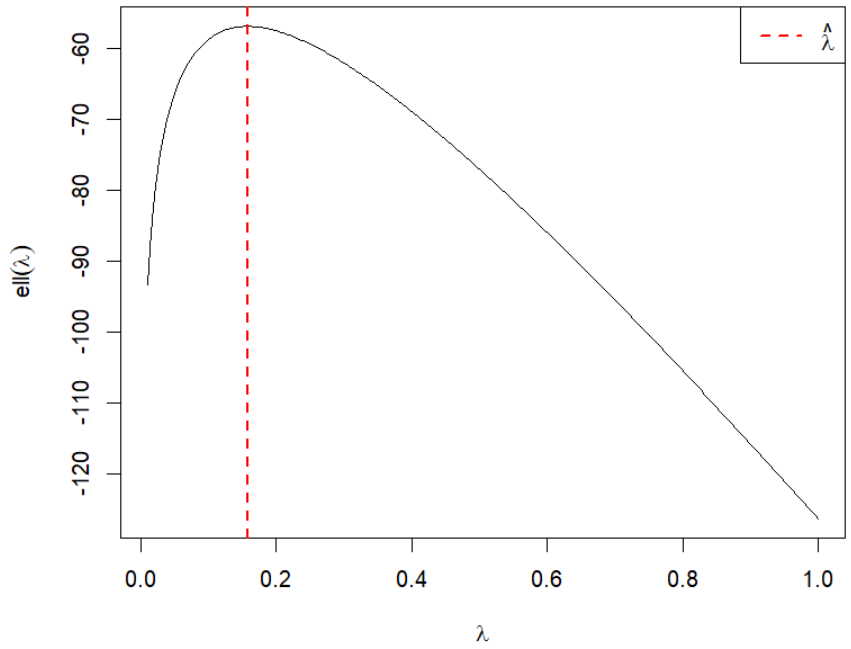

Figura 1: Gráfico da função log-verossimilhança

1.5) Tempo médio de vida e probabilidade de funcionamento.

Utilizando o parâmetro estimado $\hat{\lambda}$, obtêm-se as seguintes quantidades de interesse:

(a) Tempo médio de vida estimado

Para a distribuição exponencial, o valor esperado é dado por:

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

Substituindo λ pelo estimador de máxima verossimilhança, obtém-se o tempo médio estimado:

$$\widehat{E(X)} = \frac{1}{\hat{\lambda}} = \frac{1}{0.15832} \approx 6.32.\text{(No R: 6.3145)}$$

```r
# 1.5(a) Tempo médio de vida estimado
# Para a exponencial, a expectativa é 1/λ
expectativa_exp <- function(lambda){
  1 / lambda
}
tempomedio <- expectativa_exp(estimador)
```

Listado 7: Tempo médio de vida estimado

(b) Probabilidade de funcionamento por mais de 5 anos

A probabilidade de que um computador funcione por mais de 5 anos é dada pela função de sobrevivência da distribuição exponencial:

$$P(X > 5) = e^{-5\hat{\lambda}} = e^{-5 \cdot 0.15832} \approx 0.453. \text{(No R: 0.4530)}$$

```r
# 1.5(b) Probabilidade de o computador durar mais de 5 anos, ou seja, P(X > 5)
prob5b <- exp(-estimador * 5)
```

Listado 8: Probabilidade de funcionamento por mais de 5 anos

1.6) Propriedade da falta de memória.

A distribuição exponencial possui a propriedade da falta de memória, segundo a qual a probabilidade de funcionamento por um tempo adicional não depende do tempo já decorrido. Matematicamente, essa propriedade é expressa por:

$$P(X > s + t \mid X > s) = P(X > t).$$

Em termos práticos, isso significa que um computador que já funcionou por vários anos tem a mesma probabilidade de continuar funcionando por mais um determinado período que um computador novo.

Apesar de essa hipótese ser adequada para modelar falhas aleatórias, ela pode não ser totalmente realista para computadores, uma vez que fatores como desgaste físico e obsolescência tendem a aumentar a probabilidade de falha ao longo do tempo.

2 Questão 2

O conjunto de dados de `penguins`, na biblioteca `palmerpenguins`[1] do R, contém medidas para as três espécies de pinguins (figura 2): 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.

[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`

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.

Solução da questão

2.1) Gráfico de dispersão entre x e y.

Para gerar o gráfico da Figura 3, o qual representa a relação entre a massa corporal e o comprimento do bico dos pinguins do banco de dados fornecido, foi utilizado o código em R abaixo.

```r
# ----------- QUESTÃO 2 ----------- #
penguins_data <-na.omit(penguins) # desconsiderando os dados faltantes

# 2.1) Gráfico de dispersão entre x (massa) and y (comprimento do bico)
x <- penguins_data$body_mass_g
```

```r
y <- penguins_data$bill_length_mm

ggplot(penguins_data, aes(x = body_mass_g, y = bill_length_mm)) +
  geom_point() +
  labs(title = "",
       x = "Massa Corporal (g)",
       y = "Comprimento do Bico (mm)")
```

Listado 9: Gráfico de dispersão gerado.

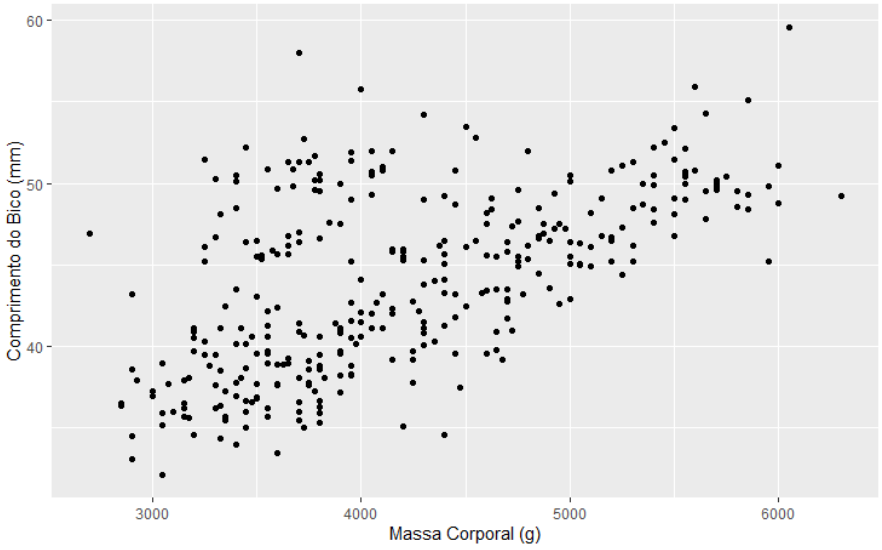

Figura 3: Gráfico de dispersão entre a massa corporal e o comprimento do bico

◦ Análise do gráfico

A partir da interpretação do gráfico de dispersão gerado, pode-se observar que existe certa plausibilidade no estabelecimento de uma relação linear entre a massa corporal e o comprimento do bico na amostra de pinguins analisada. Isso acontece porque percebe-se uma visível concentração dos pontos em uma área situada em torno de uma linha imaginária traçada a partir de uma região muito próxima à origem. Apesar disso, também podem-se visualizar vários dados registrados distantes dessa área, o que demonstra o caráter aproximativo desse modelo.

2.2) Reta de regressão linear

Para definir os parâmetros da reta de regressão utilizando método dos mínimos quadrados (MMQ), foram criadas duas funções em R (`beta0` e `beta1`), as quais aplicam as fórmulas utilizadas para o cálculo dos coeficientes mostradas abaixo.

Equação da reta:

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

Coeficientes de mínimos quadrados:

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

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

Em seguida, para realizar a verificação exigida, os valores gerados pelo MMQ foram comparados com os resultantes do comando `lm()`, função central na Linguagem R utilizada para criar modelos lineares e permitir análises estatísticas, já que fornece dados como estimativas e erro padrão. Para extrair apenas os coeficientes das informaçõe geradas, utilizou-se a função `coef()`, como mostrado no código abaixo.

```r
# 2.2) Regressão linear

#Método dos mínimos quadrados
dados <- data.frame(x, y)
n <- length(x)
beta1 <- (n * sum(x*y) - sum(x) * sum(y)) / (n * sum(x^2) - (sum(x))^2)
beta0 <- mean(y) - beta1 * mean(x)

mmq <- c(beta0, beta1)

# Função Lm
modelo <- lm(y ~ x, data = dados)
funcao_lm <- as.numeric(coef(modelo))

#Comparação dos resultados
tabela <- data.frame(
  Metodo = c("MMQ", "Comando lm()"),
  Intercepto = c(mmq[1], funcao_lm[1]),
  Inclinação = c(mmq[2], funcao_lm[2])
)

print(tabela)

summary(modelo) #mostra informações geradas por lm()
```

Listado 10: Cálculo e verificação da reta de regressão

```
        Metodo Intercepto Inclinação
1          MMQ   27.15072  0.00400329
2 Comando lm()   27.15072  0.00400329

Call:
lm(formula = y ~ x, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-10.1652  -3.0664  -0.7672   2.2356  16.0371

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.715e+01  1.292e+00   21.02   <2e-16 ***
x           4.003e-03  3.016e-04   13.28   <2e-16 ***
---

Residual standard error: 4.424 on 331 degrees of freedom
Multiple R-squared:  0.3475,    Adjusted R-squared:  0.3455
F-statistic: 176.2 on 1 and 331 DF,  p-value: < 2.2e-16
```

Listado 11: Saída do código no console

Percebe-se, então, pela tabela gerada que os resultados de ambos os métodos são iguais, o que pode ser explicado pelo fato de que a função `lm()` também usa o MMQ para o cálculo dos coeficientes da reta de regressão linear.

Além disso, a partir da observação dos outros dados fornecidos por esse comando, nota-se que o teste F do modelo (`F-statistic`), o qual testa a utilidade da sua aplicação, fornece um `p-value` associado muito próximo de zero, o que corrobora a hipótese da influência linear de x em y e permite considerar o modelo estatisticamente significativo.

Para analisar tal interpretação, foi gerado pelo código abaixo o gráfico de dispersão entre as grandezas com a reta de regressão linear (Figura 4).

```
#Gráfico com reta de regressão
ggplot(penguins_data, aes(x = body_mass_g, y = bill_length_mm)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "",
       x = "Massa Corporal (g)",
       y = "Comprimento do Bico (mm)")
```

Listado 12: Código utilizado para gerar o gráfico.

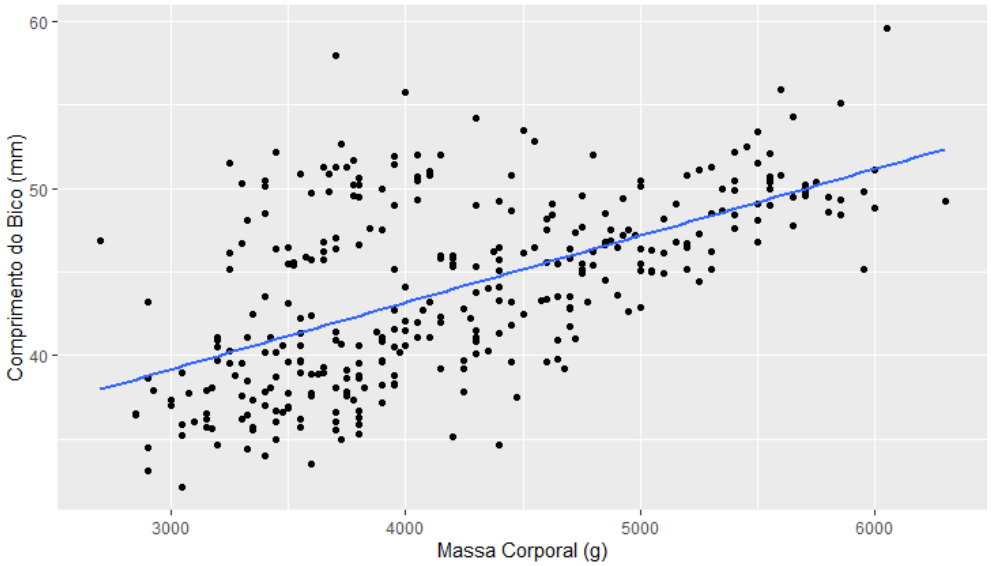

Figura 4: Gráfico de dispersão entre a massa corporal e o comprimento do bico com reta de regressão linear

◦ **Análise do gráfico**

Observa-se na figura que, apesar da interpretação visual do gráfico feita no item anterior e das informações fornecidas por `lm()` indicando a existência de uma associação linear estatisticamente significativa entre massa corporal e comprimento do bico, a reta desenhada parece levemente deslocada da nuvem de pontos, o que pode ser explicado pelo fato do modelo não considerar a heterogeneidado gerada por outros fatores, como as diferentes espécies de pinguins.

2.3) Cálculo dos resíduos da regressão, da raiz do erro quadrático médio e o coeficiente de determinação R^2.

◦ Resíduos da regressão linear

Para calcular os resíduos da regressão linear, foi extraído o parâmetro `residuals` do modelo linear gerado pelo comando `lm()`, o qual calcula o resíduo em cada ponto utilizando a fórmula:

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

na qual, y_i representa o valor observado da variável e $\hat{y}_i$ o valor estimado pelo modelo de regressão.

A seguir o código utilizado para obter os resíduos e representá-los graficamente na Figura 5.

```r
# 2.3) 1.Resíduos
resíduos <- modelo$residuals

#Gráfico para visualização
plot(penguins_data$body_mass_g, resíduos,
     xlab = "Massa Corporal (g)",
     ylab = "Resíduos",
     main = "",
     pch = 16)

abline(h = 0, lwd = 2)
```

Listado 13: Código para cálculo e representação gráfica dos resíduos

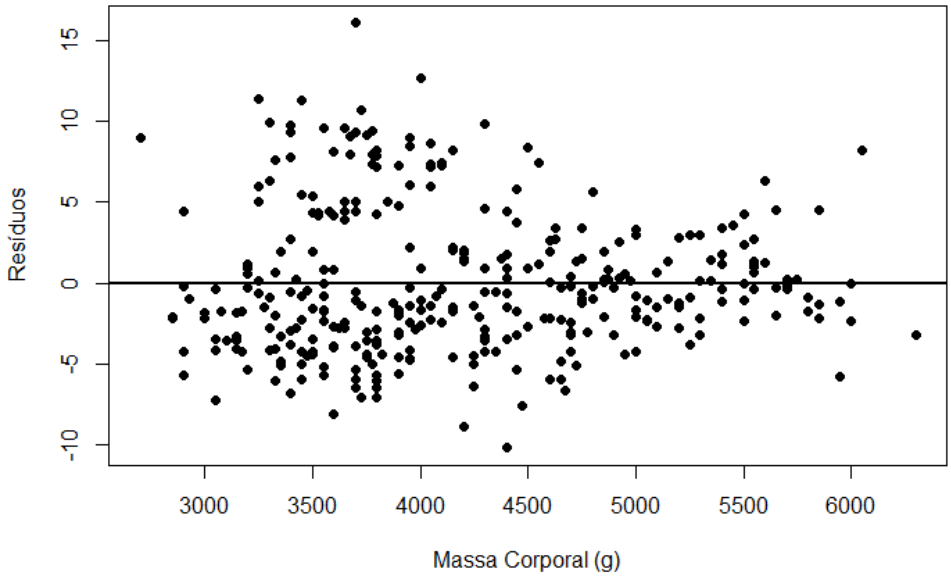

Figura 5: Gráfico da relação entre resíduos da regressão e a massa corporal

◦ Análise do gráfico

Observa-se na figura que os resíduos não se distribuem aleatoriamente em torno do zero, ou seja, exigem certos padrões de concentração dos pontos, como a maior concentração de resíduos para valores maiores de massa em relação aos menores. Essa percepção evidencia que os erros do modelo não tratam-se apenas de ruídos, mas sim de critérios desconsiderados, como a espécie dos pinguins, que causam heterogeneidade nos dados e, consequentemente, resultam nessa relação visível entre x e a variância para cada ponto.

- **Raiz do erro quadrático médio (RSME) e coeficiente de determinação R^2**

Sabendo que o RSME representa o erro médio das previsões do modelo, a fórmula utilizada para o seu cálculo é:

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

- n é o número total de observações;
- y_i é o valor observado da variável resposta na i-ésima observação;
- $\hat{y}_i$ é o valor ajustado (previsto) pelo modelo de regressão;
- $e_i = y_i - \hat{y}_i$ é o resíduo associado à i-ésima observação.

Enquanto isso, o R^2 representa a proporção da variância de y explicada pelo modelo e, no caso da regressão linear simples, pode ser calculado como:

$$R^2 = (\text{corr}(x, y))^2$$

Assim, para o cálculo desses dados, pode-se utilizar, para o RSME, a média dos quadrados dos resíduos obtidos anteriormente e, para o R^2, o próprio `r.squared` do modelo fornecido pelo `sumary(modelo)` no item anterior, como mostra o código abaixo.

```r
#Cálculo do RSME
cat('Raiz do erro quadrático médio:', "\n")
rmse
rmse <- sqrt(mean(resíduos^2))

#Cálculo do R^2
cat("Coeficiente de determinação:", "\n")
summary(modelo)$r.squared
```

Listado 14: Código em R para o cálculo do RSME e de R^2

```
Raiz do erro quadrático médio:
[1] 4.410974
Coeficiente de determinação:
[1] 0.3474526
```

Listado 15: Saída do código no console

A partir dos resultados, percebe-se que as previsões para o comprimento do bico dos pinguins diferem dos valores observados em aproximadamente $4.4mm$, o que se trata de valor pequeno em comparação a faixa de tamanhos observada. Além disso, pela análise do R^2, sabe-se que, aproximadamente, 35% da variabilidade observada no comprimento do bico pode ser atribuída à massa corporal, enquanto o restante está relacionada a outros fatores, como a espécie, por exemplo. Dessa froma, esses dados permitem concluir que apesar de apresentarem um baixo erro absoluto, a variabilidade moderada não pode ser explicada pelo modelo linear.

2.4) Introduzindo uma observação artifical extrema na massa corporal

Para observar a influência de uma observação artificial extrema nos resultados da regressão, foi escolhida a introdução de um penguim macho com uma massa corporal significativamente maior que a faixa de valores registrada e a análise dos novos dados gerados em comparação ao modelo anterior, como mostrado no código abaixo.

```
# 2.4) Introdução de um outlier

# Criando um ponto extremo
penguins_outlier <- rbind(
  penguins_data ,
  data.frame(
    species = "Adelie",
    island = "Torgersen",
    bill_length_mm = 45,   # valor típico
    bill_depth_mm = 19,
    flipper_length_mm = 190,
    body_mass_g = 8000,    # valor muito alto
    sex = "male",
    year = 2007
  )
)

modelo_outlier <- lm(bill_length_mm ~ body_mass_g,
                     data = penguins_outlier)

summary(modelo_outlier)

# novos valores para RSME e R^2
rmse_outlier <- sqrt(mean(residuals(modelo_outlier)^2))
r2_outlier <- summary(modelo_outlier)$r.squared
r2_original <- summary(modelo)$r.squared

#comparando coeficientes
coef_original <- coef(modelo)
coef_outlier <- coef(modelo_outlier)

tabela_comparacao <- data.frame(
  Modelo = c("Original", "Com outlier"),
  Intercepto = c(coef_original[1], coef_outlier[1]),
  Inclinação = c(coef_original[2], coef_outlier[2]),
  RMSE = c(rmse, rmse_outlier),
  R2 = c(r2_original, r2_outlier)
)
tabela_comparacao
```

Listado 16: Código com a introdução do outlier

```
                Modelo Intercepto  Inclinação     RMSE        R2
x              Original   27.15072 0.004003290 4.410974 0.3474526
body_mass_g Com outlier   28.09326 0.003769793 4.467765 0.3285996
```

Listado 17: Saída do código no console

A partir dessa comparação, percebe-se que a adição do outlier provocou uma redução da inclinação e aumento do intercepto, o que evidencia que a reta ajustou-se para tentar acomodar a nova observação, mostrando que o outlier puxou a regressão. Além disso, também observou-se um aumento no RSME e uma redução no R^2, o que indica uma piora no ajuste do modelo, ou seja, que a adição desse dado extremo reduziu a significância estatística do modelo linear.

Para comparar visualmente a influência dessa observação, foi gerado o gráfico da Figura 6, de modo a evidenciar as diferenças geradas na reta de regressão linear correspondente ao modelo.

```r
plot(penguins_outlier$body_mass_g,
     penguins_outlier$bill_length_mm,
     xlab = "Massa Corporal (g)",
     ylab = "Comprimento do Bico (mm)",
     main = "",
     pch = 16,
     col = "grey50")

# Reta do modelo ORIGINAL
abline(modelo,
       col = "blue",
       lwd = 2)

# Reta do modelo COM OUTLIER
abline(modelo_outlier,
       col = "red",
       lwd = 2)

# Legenda
legend("topleft",
       legend = c("Modelo original", "Modelo com outlier"),
       col = c("blue", "red"),
       lwd = 2,
       bty = "n")
```

Listado 18: Código para gráfico de comparação

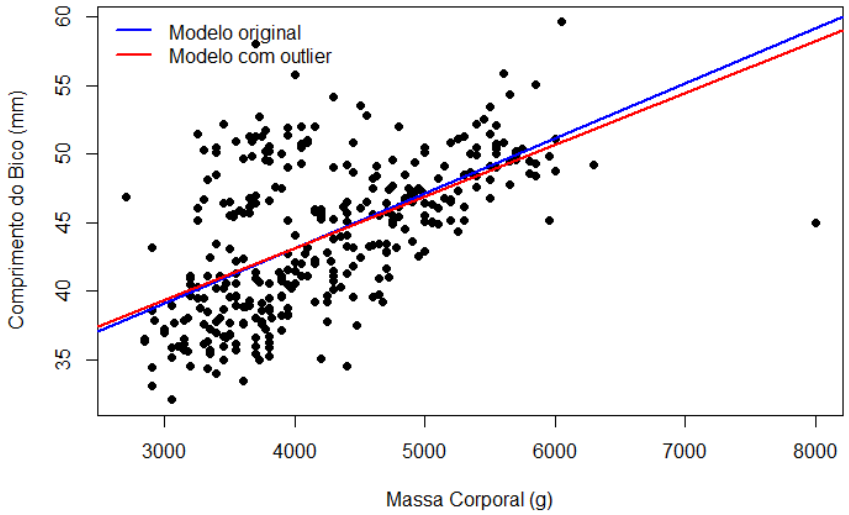

Figura 6: Gráfico de comparação entre as retas de regressão do modelo com e sem o outlier

◦ **Análise do gráfico**

A partir da análise do gráfico, é possível perceber as mudanças dos valores na tabela de comparação, como o leve deslocamento para cima e a redução na inclinação da reta. Apesar de, visualmente, as diferenças entre a reta do modelo original e a do modelo com o outlier (representado pelo ponto mais distante situado em 8000g) não serem significativas, deve-se considerar que elas foram resultantes da adição de apenas uma observação extrema em um grupo de muitos pontos, o que evidencia o potencial da influência dos outliers no modelo de regressão linear.