ESTATÍSTICA PARA ENGENHARIA

HOMEWORK 3

Alunos:
JOAO VICTOR MARQUES FALCÃO, 567357
LUCAS MARTINS MENEZES, 565788

1 Tempo de vida de um computador

1.1 Função densidade de probabilidade

A distribuição exponencial é uma distribuição contínua frequentemente usada para modelar o tempo até a ocorrência de um evento, como o tempo de vida de um computador ou a duração de um processo de espera. Ela tem a propriedade interessante de ser sem memória, o que significa que a probabilidade de um evento ocorrer no futuro é independente do tempo já decorrido.

A função de densidade de probabilidade (pdf) da distribuição exponencial com parâmetro λ é dada por:

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

Explicação dos componentes: - x: Representa o valor da variável aleatória, neste caso o tempo de vida de um computador. A variável x é sempre não-negativa, pois não faz sentido ter um tempo de vida negativo. - λ: É o parâmetro da taxa de falha, também chamado de taxa de evento. Ele controla a intensidade da falha: se λ for grande, isso significa que o evento (a falha) ocorre mais rapidamente. Se λ for pequeno, o evento ocorre mais lentamente. Em termos de unidades, λ tem a unidade de inverso do tempo (no caso, anos^{-1}). - $e^{-\lambda x}$: Essa parte da fórmula descreve a função exponencial que determina como a probabilidade diminui com o tempo. Quanto maior o tempo x, menor a probabilidade de o evento ocorrer nesse tempo.

O fato de a densidade ser zero para $x < 0$ reflete a natureza não negativa de x, já que estamos lidando com o tempo de vida de algo, que não pode ser negativo.

Propriedade de Normalização: Uma distribuição de probabilidade precisa atender à condição de que a soma (ou integral) de todas as probabilidades seja igual a 1. Ou seja, se integramos a função $f_X(x;\lambda)$ sobre todo o intervalo de x (de $-\infty$ a ∞), o resultado deve ser 1. No caso da distribuição exponencial:

$$\int_0^\infty \lambda e^{-\lambda x}\, dx = 1.$$

Essa integral é simples de calcular, pois é uma forma padrão, e o resultado será 1, garantindo que a probabilidade total seja 100%. Isso é importante, pois sem essa propriedade, não poderíamos usá-la para modelar probabilidades.

1.2 Função de Verossimilhança

Considerando que temos uma amostra de n observações $X_1, X_2, \ldots, X_n$ de uma variável aleatória que segue uma distribuição exponencial $X \sim \text{Exp}(\lambda)$, a função de verossimilhança $L(\lambda)$ é o produto das densidades de probabilidade das observações individuais.

Cada X_i tem a densidade $f(x_i;\lambda) = \lambda e^{-\lambda x_i}$, então a função de verossimilhança é dada por:

$$L(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \left[\lambda e^{-\lambda x_i}\right].$$

Ou seja, a função de verossimilhança é simplesmente o produto de todas as probabilidades individuais associadas a cada uma das observações $x_1, x_2, \ldots, x_n$.

Ao simplificar, temos:

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

Aqui: - λ^n vem do produto das n densidades exponenciais, cada uma com o fator λ. - $e^{-\lambda \sum_{i=1}^{n} x_i}$ é o produto das exponenciais, que se tornam somas quando multiplicadas.

A função de verossimilhança é uma função de λ que expressa a "probabilidade"de observarmos os dados $X_1, X_2, \ldots, X_n$, dado o parâmetro λ.

A função log-verossimilhança $\ell(\lambda)$ é simplesmente o logaritmo da função de verossimilhança. Ela é útil porque transforma o produto da verossimilhança em uma soma, o que facilita os cálculos, principalmente ao derivarmos para encontrar o estimador de máxima verossimilhança (MLE). Temos:

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

1.3 Estimador de Máxima Verossimilhança (MLE)

O estimador de máxima verossimilhança (MLE) $\hat{\lambda}$ é o valor de λ que maximiza a função log-verossimilhança $\ell(\lambda)$. Para encontrar o valor de λ que maximiza $\ell(\lambda)$, fazemos a derivada de $\ell(\lambda)$ em relação a λ, igualamos a derivada a zero e resolvemos para λ.

A derivada de $\ell(\lambda)$ em relação a λ é:

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

A equação é igualada a zero para maximizar a função:

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0.$$

Resolvendo para λ:

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

Ou, como $\bar{x}$ é a média amostral ($\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$), podemos escrever:

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

Esse é o estimador de máxima verossimilhança. Ele nos dá o valor de λ que torna os dados observados mais prováveis.

1.4 Gráfico da Função Log-Verossimilhança

A função log-verossimilhança $\ell(\lambda)$ é uma função que depende de λ. Ao variar λ, obtemos diferentes valores de $\ell(\lambda)$. O gráfico dessa função mostra como $\ell(\lambda)$ se comporta para diferentes valores de λ, permitindo visualizar onde a função atinge seu máximo (onde $\hat{\lambda}$ está localizado).

Podemos obter uma conclusão mais visual com o gráfico gerado a partir do seguinte código em R:

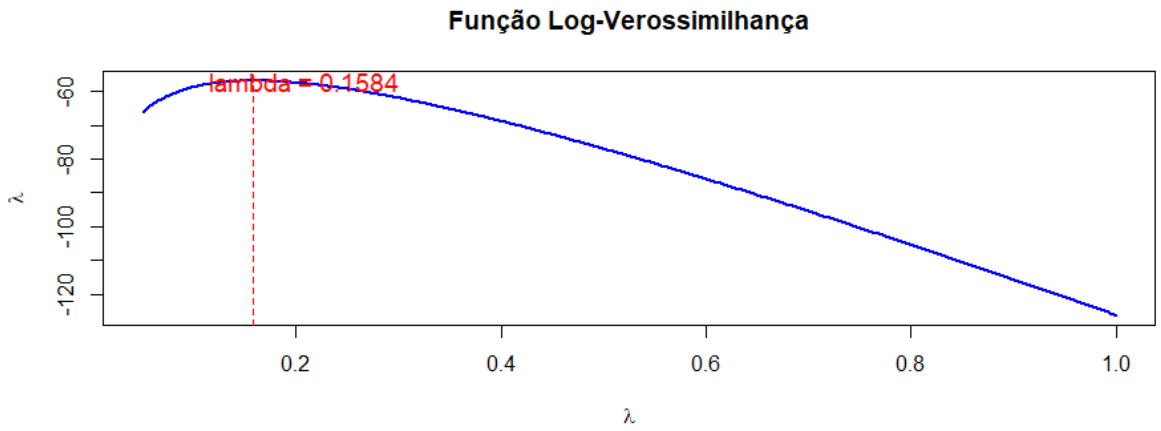

Figura 1: Gráfioco da Função Log-Verossimilhança

```r
# dados
n <- 20  # n mero de observacoes
sum_x <- 126.29  # soma dos valores dos dados

# Intervalo para lambda
lambda_values <- seq(0.05, 1.0, length.out = 500)

# verossimilhanca
log_likelihood <- n * log(lambda_values) - lambda_values * sum_x

# mle
lambda_hat <- 1 / (sum_x / n)

# grafico
plot(lambda_values, log_likelihood, type = "l", col = "blue",
     lwd = 2, xlab = expression(lambda), ylab = expression(lambda),
     main = "Fun  o Log-Verossimilhan a")


abline(v = lambda_hat, col = "red", lty = 2)


text(lambda_hat + 0.05, max(log_likelihood),
     labels = paste("lambda =", round(lambda_hat, 4)),
     col = "red", cex = 1.2)
```

Com ele, obtemos o seguinte gráfico:

1.5 Usando o parâmetro

(a) Tempo Médio de Vida Estimado

A distribuição exponencial tem uma característica importante: o tempo médio de vida $E[X]$ de uma variável X que segue a distribuição exponencial com parâmetro λ é dado por:

$$E[X] = \frac{1}{\lambda}.$$

Com o estimador $\hat{\lambda} \approx 0.1584$, o tempo médio de vida estimado do computador é:

$$\hat{E}[X] = \frac{1}{\hat{\lambda}} = \frac{1}{0.1584} \approx 6.31 \text{ anos.}$$

Ou seja, com base nos dados observados, estima-se que o tempo médio de vida do computador seja de aproximadamente 6.31 anos.

(b) Probabilidade de Durar Mais de 5 Anos

A probabilidade de que um computador dure mais de 5 anos, dado que a variável X segue uma distribuição exponencial com parâmetro λ, é dada pela **função de sobrevivência** da distribuição exponencial:

$$P(X > 5) = e^{-\lambda \cdot 5}.$$

Substituindo $\hat{\lambda} = 0.1584$:

$$P(X > 5) = e^{-0.1584 \cdot 5} \approx e^{-0.792} \approx 0.453.$$

Portanto, a probabilidade de que um computador dure mais de 5 anos é aproximadamente 45.3%.

1.6 Propriedade da Falta de Memória

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) Explicação da propriedade

A *propriedade da falta de memória* significa que, dado que o computador já está em funcionamento por um certo tempo, a **probabilidade de falha no futuro** não depende de quanto tempo ele já está em operação. Ou seja, o comportamento futuro do computador não é afetado pelo seu tempo de operação anterior.

Para formalizar, para uma variável aleatória X com distribuição exponencial (representando o tempo de vida do computador), temos a seguinte propriedade:

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

onde: - $P(X > t + s \mid X > t)$ é a probabilidade de o computador funcionar por mais s anos, dado que ele já funcionou por t anos. - $P(X > s)$ é simplesmente a probabilidade de o computador funcionar por mais s anos, sem levar em conta quanto tempo ele já funcionou.

Esta propriedade mostra que, independentemente de quanto tempo o computador já funcionou, a probabilidade de falhar no futuro é sempre a mesma.

(b) Discussão sobre a razoabilidade dessa suposição para modelar o tempo de vida de computadores

A suposição de que o tempo de vida de um computador segue uma distribuição exponencial e possui a *propriedade da falta de memória* pode ser razoável em certos contextos, mas nem sempre é adequada para modelar sistemas complexos como computadores.

Razoabilidade: Para muitos sistemas simples, a distribuição exponencial pode ser uma boa aproximação, especialmente se as falhas ocorrerem de maneira independente e com uma taxa constante ao longo do tempo. Em alguns modelos de falha de computadores (como componentes eletrônicos com falhas aleatórias), essa suposição pode ser aceitável.

Limitações: No entanto, em sistemas mais complexos, a falha de um computador pode depender de fatores como a idade do componente, condições de uso, e manutenção. A distribuição exponencial ignora esses efeitos, o que pode não ser realista, pois muitas falhas de computadores tendem a ocorrer após um certo período de uso, ou devido ao desgaste acumulado. Em situações como essas, distribuições como a **Weibull** ou **Gamma** podem ser mais apropriadas, pois essas distribuições permitem modelar a taxa de falha que varia ao longo do tempo.

2 Regressão Linear: Biometria e Conservação dos Pinguins de Palmer

A análise estatística de características morfológicas em populações biológicas, área conhecida como biometria, é uma ferramenta essencial para a conservação e o monitoramento ambiental. No caso dos pinguins das espécies Adélia, Gentoo e Barbicha (Chinstrap) do arquipélago Palmer, o estudo das relações alométricas permite que pesquisadores compreendam como a massa corporal reflete o desenvolvimento de estruturas específicas, como o comprimento do bico.

Do ponto de vista prático e aplicado à vida real, a modelagem dessas variáveis justifica-se por dois pilares fundamentais:

- **Eficiência de Campo:** Em expedições na Antártica, mensurar o comprimento do bico com precisão milimétrica pode ser complexo sob condições climáticas adversas. Um modelo de regressão robusto permite estimar essa dimensão a partir da massa corporal — uma medida de captura mais simples e rápida — otimizando o tempo de coleta e reduzindo o estresse do animal.

- **Monitoramento Ecossistêmico:** Alterações nos coeficientes de regressão (β_1) ou aumentos significativos nos resíduos ao longo dos anos podem atuar como indicadores precoces de mudanças climáticas ou desequilíbrios na cadeia alimentar, impactando a saúde nutricional das colônias.

Nesta seção, utiliza-se o conjunto de dados *palmerpenguins* para parametrizar essa relação, transformando observações biológicas em um sistema linear preditivo sujeito a diagnósticos de precisão e robustez.

A estrutura do conjunto de dados é consolidada em oito variáveis fundamentais, as quais permitem a análise multidimensional da fauna observada. A Tabela 1 detalha a natureza técnica de cada coluna e identifica as variáveis que comporão o núcleo do modelo de regressão linear simples.

Tabela 1: Dicionário de variáveis do *dataset* `penguins_data.csv`.

Variável	Descrição Técnica	Papel no Modelo
`species`	Espécie do pinguim (*Adelie*, *Chinstrap* ou *Gentoo*).	Categórica
`island`	Local de coleta no arquipélago Palmer (*Biscoe*, *Dream* ou *Torgersen*).	Contexto
`bill_length_mm`	Comprimento do bico em milímetros (mm).	Dependente (y)
`bill_depth_mm`	Profundidade vertical do bico em milímetros (mm).	Controle
`flipper_length_mm`	Comprimento da nadadeira em milímetros (mm).	Controle
`body_mass_g`	Massa corporal total em gramas (g).	Independente (x)
`sex`	Sexo do indivíduo (*male* ou *female*).	Categórica
`year`	Ano cronológico em que a medição foi realizada.	Temporal

2.1 Análise Exploratória e Plausibilidade Linear

A etapa inicial de qualquer modelagem preditiva reside na Análise Exploratória de Dados (AED), que atua como um filtro crítico para a validação das premissas estatísticas fundamentais antes da aplicação de métodos de inferência. No contexto do estudo dos pinguins do arquipélago Palmer, a inspeção visual por meio de um gráfico de dispersão (*scatterplot*) é o procedimento técnico que permite diagnosticar a natureza da associação entre a variável independente massa corporal (x) e a variável dependente comprimento do bico (y), identificando antecipadamente padrões de correlação ou potenciais desvios de linearidade que poderiam invalidar o uso de estimadores clássicos.

Para a visualização desta relação, implementou-se o seguinte código no ambiente `R`, incorporando uma grade de coordenadas para facilitar a associação entre as grandezas físicas observadas:

```
# Plotagem do gr fico de dispers o para an lise de
   plausibilidade
plot(penguins_data$body_mass_g, penguins_data$bill_length_mm,
     pch = 16, col = rgb(0, 0, 1, 0.5),
     main = "Massa␣Corporal␣vs.␣Comprimento␣do␣Bico",
     xlab = "Massa␣Corporal␣(g)",
     ylab = "Comprimento␣do␣Bico␣(mm)")

# Adi   o de grade para aux lio na leitura de tend ncias
grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted")
```

A partir da observação da Figura 2, constata-se uma tendência ascendente nítida, revelando uma correlação positiva entre o peso do espécime e suas dimensões cranianas. Embora o conjunto de dados apresente uma dispersão considerável — fenômeno intrínseco à variabilidade biológica entre as três espécies e ao dimorfismo sexual presente nas colônias —, a nuvem de pontos distribui-se de forma consistente ao longo de um eixo retilíneo. A inexistência de curvaturas acentuadas ou padrões em "S"indica que modelos de maior complexidade, como regressões polinomiais ou transformações logarítmicas, não são necessários para uma representação fiel da tendência média da população.

A confirmação da plausibilidade linear nesta fase possui implicações práticas significativas para o monitoramento ambiental e a eficiência de campo. Do ponto de vista técnico, ela justifica o emprego do Método dos Mínimos Quadrados Ordinários (OLS) como uma ferramenta de predição robusta e confiável. Na prática, essa observação assegura que o uso de pesagens rápidas em expedições antárticas é um método estatisticamente sólido para inferir o desenvolvimento morfológico dos animais, permitindo a coleta de dados em larga escala com baixo impacto ao bem-estar dos pinguins e alta precisão nos resultados estimados.

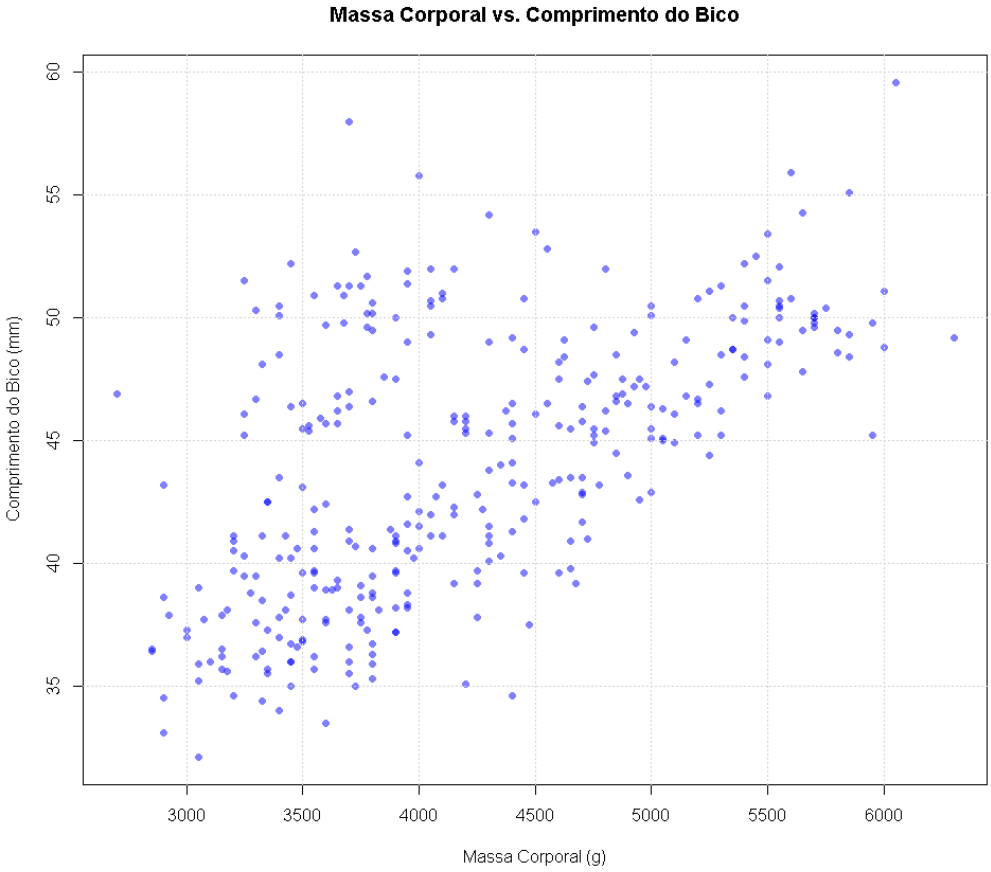

Figura 2: Gráfico de dispersão evidenciando a tendência linear positiva entre a massa corporal e o bico.

2.2 Estimação de Parâmetros via Mínimos Quadrados (OLS)

Após a validação visual da tendência linear, o próximo estágio técnico consiste na determinação analítica da reta que melhor sintetiza a relação entre a massa e a morfologia craniana dos espécimes. Para este fim, utiliza-se o Método dos Mínimos Quadrados Ordinários (OLS), um procedimento de otimização cujo objetivo primordial é minimizar a soma dos quadrados das diferenças entre os valores observados e as predições do modelo. Sob o ponto de vista estatístico, este método busca o estimador que apresente a menor variância possível entre todos os estimadores lineares não enviesados, conferindo robustez matemática às inferências realizadas sobre a população do arquipélago Palmer [**?**].

A formalização deste modelo exige a estimação de dois coeficientes fundamentais: o coeficiente angular ($\hat{\beta}_1$) e o intercepto ($\hat{\beta}_0$). As equações analíticas que regem esses estimadores baseiam-se na covariância entre as variáveis e na variância da variável independente, conforme as expressões abaixo:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{e} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

A aplicação destas fórmulas ao conjunto de dados resultou em uma inclinação $\hat{\beta}_1 \approx 0{,}0040$. Em termos práticos, este valor quantifica a taxa de variação morfológica: para cada grama adicional de massa corporal, o comprimento do bico do pinguim tende a crescer aproximadamente 0,004 mm. Esta constante alométrica é de vital importância para a engenharia biológica, pois permite caracterizar o padrão de crescimento das espécies e identificar desvios que possam indicar condições anômalas de desenvolvimento na colônia.

O intercepto calculado, $\hat{\beta}_0 \approx 27{,}1507$ mm, representa o ponto de interseção da reta com o eixo das ordenadas. Embora, sob a ótica da biologia, um indivíduo de massa zero seja uma impossibilidade física, o intercepto é matematicamente indispensável para o posicionamento correto da reta no plano cartesiano, garantindo que o modelo passe pelo centro de gravidade amostral $(\bar{x}, \bar{y})$. Assim, a equação preditiva consolida-se como $\hat{y} = 27{,}1507 + 0{,}0040x$, fornecendo uma ferramenta de estimativa rápida e confiável para o monitoramento de campo.

A validação destes cálculos foi efetuada através da linguagem `R`, onde a função `lm()` foi utilizada para ajustar o modelo e extrair o sumário estatístico completo, permitindo a posterior sobreposição da reta de regressão aos dados originais para confirmação do ajuste.

```r
# Ajuste do modelo de regress o linear simples
modelo_linear <- lm(bill_length_mm ~ body_mass_g, data = penguins_
    data)

# Exibi  o dos coeficientes estimados e estat sticas de teste
summary(modelo_linear)

# Inclus o da reta de regress o (cor vermelha) sobre o gr fico
    de dispers o
abline(modelo_linear, col = "red", lwd = 2)
```

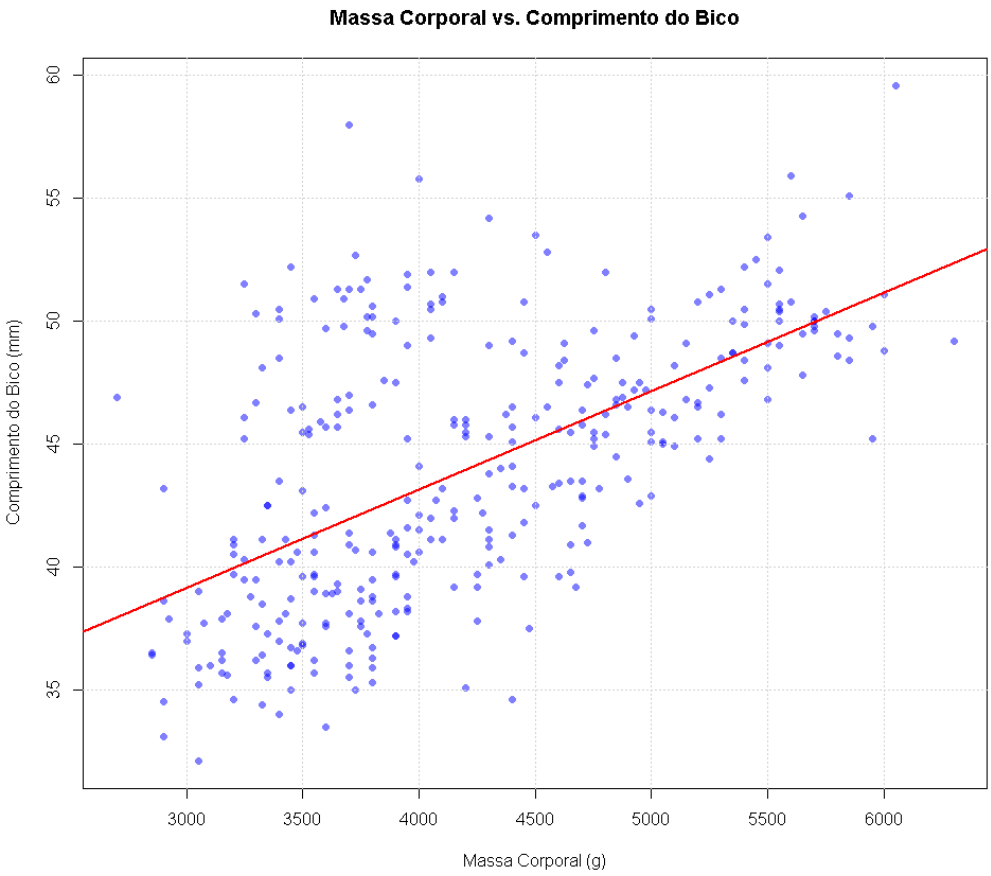

Figura 3: Reta de regressão linear (OLS) ajustada evidenciando a tendência central do crescimento do bico em função da massa.

2.3 Diagnóstico do Modelo: Resíduos, RMSE e Coeficiente de Determinação (R^2)

Após a definição dos parâmetros da reta de regressão, é fundamental avaliar a qualidade do ajuste e a capacidade preditiva do modelo proposto. O primeiro diagnóstico técnico consiste na análise dos resíduos (e_i), definidos como a discrepância entre o valor real observado e o valor estimado pelo modelo para cada indivíduo ($e_i = y_i - \hat{y}_i$). Em um modelo linear ideal, espera-se que os resíduos se distribuam de forma aleatória em torno de zero, não apresentando padrões sistemáticos que indiquem falhas na especificação da função ou violações de pressupostos básicos, como a homocedasticidade (variância constante dos erros).

Para visualizar o comportamento desses desvios, gerou-se o gráfico de resíduos em função dos valores preditos, incorporando uma linha de referência no eixo das ordenadas e uma grade de coordenadas para facilitar a inspeção de potenciais anomalias.

```
# Extra  o dos res duos do modelo ajustado
residuos <- residuals(modelo_linear)

# C lculo da Raiz do Erro Quadr tico M dio (RMSE)
rmse_val <- sqrt(mean(residuos^2))

# Obten  o do Coeficiente de Determina  o (R2)
r2_val <- summary(modelo_linear)$r.squared

# Representa  o gr fica dos res duos
plot(predict(modelo_linear), residuos,
     pch = 16, col = rgb(0.3, 0.3, 0.3, 0.6),
     main = "Gr fico de Res duos: Valores Preditos vs. Erros",
     xlab = "Valores Preditos (mm)",
     ylab = "Res duos (mm)")
abline(h = 0, col = "red", lwd = 2, lty = 2)
grid(col = "lightgray", lty = "dotted")

============================Retorno============================
Call:
lm(formula = bill_length_mm ~ body_mass_g, data = penguins_data)

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 ***
body_mass_g 4.003e-03  3.016e-04   13.28   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
       0.1         1

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
```

A eficácia do modelo é quantificada por duas métricas complementares: o RMSE e

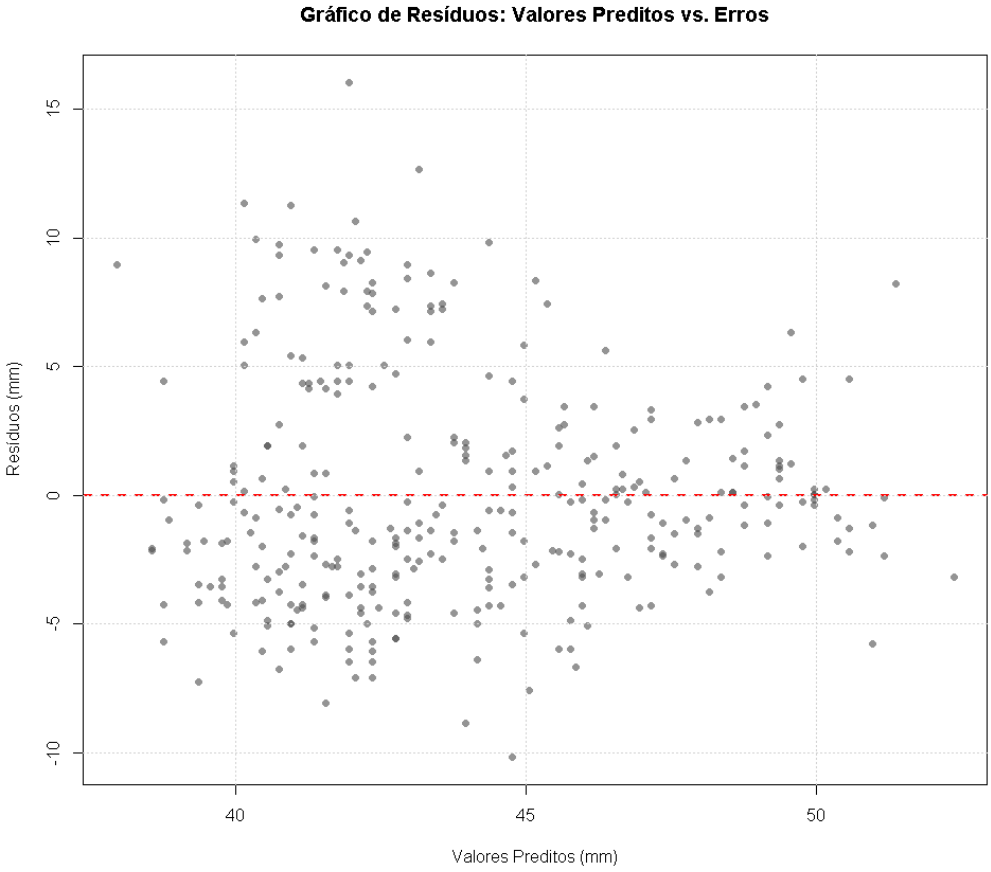

Figura 4: Distribuição dos resíduos evidenciando a dispersão dos erros de predição em relação à reta média.

o R^2. O Erro Quadrático Médio calculado foi de aproximadamente 4,41 mm, o que representa o desvio padrão típico das predições. Sob a ótica da engenharia de campo, este valor indica que, ao utilizarmos a massa corporal para estimar o comprimento do bico, devemos esperar uma incerteza média de cerca de 4,4 mm. Já o Coeficiente de Determinação ($R^2 \approx 0{,}3475$) revela que 34,75% da variabilidade total observada no comprimento do bico é explicada pela variação da massa corporal.

A interpretação destes resultados sugere que, embora a massa corporal seja um preditor estatisticamente significante, ela não captura a totalidade do fenômeno morfológico. O valor moderado do R^2 é esperado em dados biológicos complexos, onde variáveis omitidas no modelo simples — como a espécie, o sexo e o ano de coleta — exercem influência considerável sobre o desenvolvimento dos animais. Contudo, a distribuição aleatória observada na Figura 4 confirma que o modelo linear é imparcial, fornecendo uma estimativa média robusta que serve como base sólida para o monitoramento populacional no arquipélago Palmer.

2.4 Análise de Robustez e Sensibilidade a Outliers

A confiabilidade de um modelo de regressão linear em aplicações reais de engenharia e biometria depende intrinsecamente de sua estabilidade frente a observações anômalas, conhecidas como *outliers*. Uma vez que o método de Mínimos Quadrados Ordinários (OLS) baseia-se na minimização da soma dos quadrados dos resíduos, ele impõe uma penalidade quadrática aos desvios, o que confere um peso desproporcional a pontos extremos. Para avaliar essa sensibilidade no contexto do arquipélago Palmer, procedeu-se a um teste de

estresse introduzindo artificialmente uma observação espúria: alterou-se o comprimento do bico do primeiro pinguim da amostra de 39,1 mm para 80,0 mm, simulando um erro grosseiro de medição ou transcrição de dados.

O diagnóstico desta sensibilidade é realizado de forma mais contundente através da comparação direta entre o modelo íntegro e o modelo corrompido, tanto em termos de métricas estatísticas quanto por meio da sobreposição visual das retas ajustadas no plano cartesiano.

```
# Criacao de uma copia do dataset para teste de robustez
penguins_outlier <- penguins_data

# Introducao do outlier (Pinguim 1: massa 3750g | bico: 39.1mm -> 80.0mm)
penguins_outlier[1, "bill_length_mm"] <- 80.0

# Ajuste do novo modelo com os dados modificados
modelo_outlier <- lm(bill_length_mm ~ body_mass_g, data = penguins_outlier)

# Visualizacao comparativa: Dados Originais vs. Retas de Ajuste
plot(penguins_data$body_mass_g, penguins_data$bill_length_mm,
     pch = 16, col = rgb(0.5, 0.5, 0.5, 0.3),
     main = "Analise de Sensibilidade: Impacto de um Outlier",
     xlab = "Massa Corporal (g)", ylab = "Comprimento do Bico (mm)")

# Reta Original (Item 2) em Azul
abline(modelo_linear, col = "blue", lwd = 2)

# Reta com Outlier (Item 4) em Vermelho Tracejado
abline(modelo_outlier, col = "red", lwd = 2, lty = 2)

# Legenda Tecnica
legend("topleft", legend = c("Modelo Original", "Modelo com Outlier"),
       col = c("blue", "red"), lty = c(1, 2), lwd = 2, bty = "n")
grid(col = "lightgray", lty = "dotted")
```

A análise comparativa apresentada na Tabela 2 revela a vulnerabilidade do estimador OLS frente a dados espúrios. A introdução de apenas uma observação anômala em um universo de 333 amostras foi suficiente para degradar o coeficiente de determinação (R^2) em mais de 15%, além de elevar a incerteza média de predição (RMSE) em aproximadamente 10%. Conforme observado na Figura 5, a inclinação da reta sofreu uma alteração perceptível; o modelo "puxa"a reta em direção ao ponto extremo para tentar minimizar o erro quadrático total, um fenômeno conhecido como alavancagem (*leverage*), que compromete a representatividade da tendência central da população.

Em termos de conclusões práticas para a engenharia de dados e monitoramento ambiental, este experimento reforça que a qualidade da inferência é estritamente dependente da integridade da base de dados. A sensibilidade observada ressalta a importância crítica de etapas de pré-processamento, como a detecção de *outliers* e a filtragem de ruídos de sensores. Em cenários onde a ocorrência de anomalias é frequente ou inevitável, os resultados aqui obtidos justificam a transição para métodos de regressão robusta, que utilizam funções de perda menos sensíveis a valores extremos, garantindo que o sistema

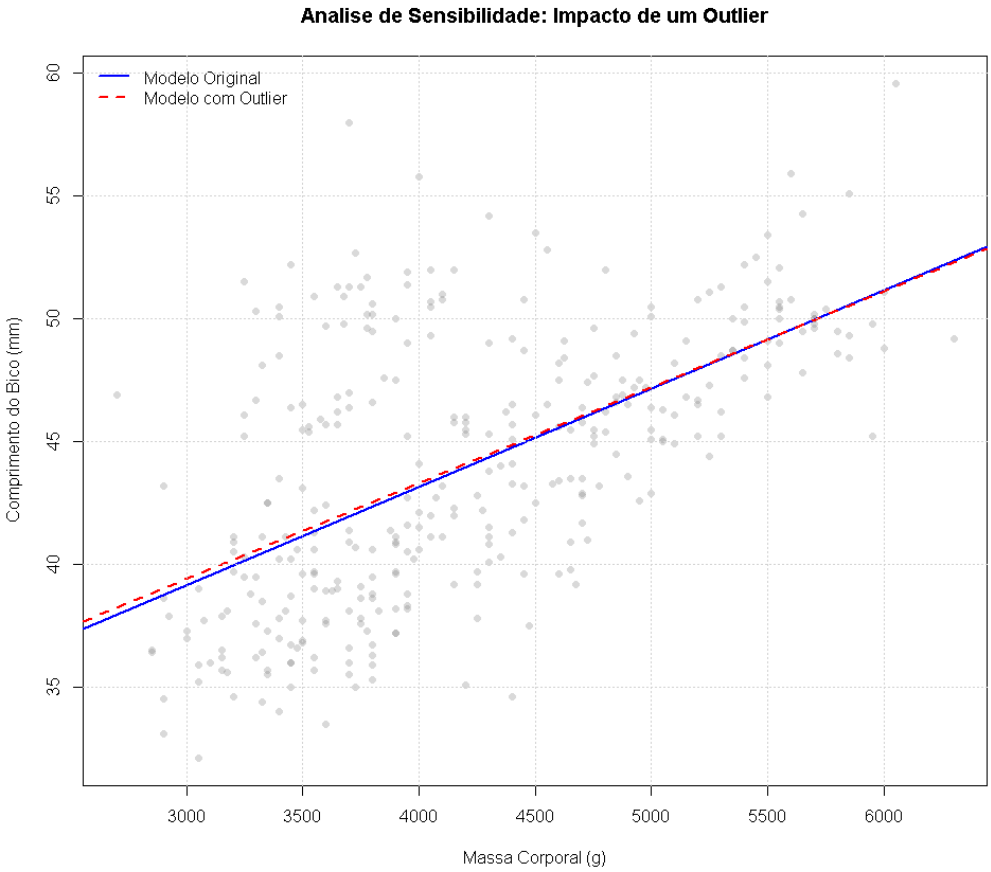

Figura 5: Comparação visual evidenciando o efeito de alavancagem provocado pelo outlier sobre a inclinação da reta.

Tabela 2: Impacto comparativo da introdução de uma observação extrema no modelo.

Métrica	**Modelo Original**	**Modelo com Outlier**	**Variação Relativa**
Intercepto ($\hat{\beta}_0$)	27,1507	27,6389	+1,80%
Inclinação ($\hat{\beta}_1$)	0,00400	0,00392	−2,00%
R^2 (Poder Explicativo)	0,3475	0,2949	−15,14%
RMSE (Incerteza Média)	4,4110	4,8691	+10,38%

de monitoramento dos pinguins de Palmer permaneça confiável mesmo diante de falhas instrumentais ou variabilidades biológicas extremas.

3 Apêndice: Código R Completo

```r
library(palmerpenguins);

penguins_data <-na.omit(penguins)

# Plotagem do gr fico de dispers o
plot(penguins_data$body_mass_g, penguins_data$bill_length_mm,
     pch = 16, col = rgb(0, 0, 1, 0.5),
     main = "Massa␣Corporal␣vs.␣Comprimento␣do␣Bico",
     xlab = "Massa␣Corporal␣(g)",
     ylab = "Comprimento␣do␣Bico␣(mm)")

# Adi   o de grade para facilitar a leitura das coordenadas
grid(nx = NULL, ny = NULL, col = "lightgray", lty = "dotted")

# Ajuste do modelo de regress o linear simples
modelo_linear <- lm(bill_length_mm ~ body_mass_g, data = penguins_
   data)

# Exibi   o dos coeficientes estimados e estat sticas de teste
summary(modelo_linear)

# Inclus o da reta de regress o (cor vermelha) sobre o gr fico
   de dispers o
abline(modelo_linear, col = "red", lwd = 2)

# Extra   o dos res duos do modelo ajustado
residuos <- residuals(modelo_linear)

# C lculo da Raiz do Erro Quadr tico M dio (RMSE)
rmse_val <- sqrt(mean(residuos^2))

# Obten   o do Coeficiente de Determina   o (R2)
r2_val <- summary(modelo_linear)$r.squared

# Representa   o gr fica dos res duos
plot(predict(modelo_linear), residuos,
     pch = 16, col = rgb(0.3, 0.3, 0.3, 0.6),
     main = "Gr fico␣de␣Res duos:␣Valores␣Preditos␣vs.␣Erros",
     xlab = "Valores␣Preditos␣(mm)",
     ylab = "Res duos␣(mm)")
abline(h = 0, col = "red", lwd = 2, lty = 2)
grid(col = "lightgray", lty = "dotted")

# Criacao de uma copia do dataset para teste de robustez
penguins_outlier <- penguins_data

```

```r
# Introducao do outlier (Pinguim 1: massa 3750g | bico: 39.1mm -> 80.0mm)
penguins_outlier[1, "bill_length_mm"] <- 80.0

# Ajuste do novo modelo com os dados modificados
modelo_outlier <- lm(bill_length_mm ~ body_mass_g, data = penguins_outlier)

# Visualizacao comparativa: Dados Originais vs. Retas de Ajuste
plot(penguins_data$body_mass_g, penguins_data$bill_length_mm,
     pch = 16, col = rgb(0.5, 0.5, 0.5, 0.3),
     main = "Analise de Sensibilidade: Impacto de um Outlier",
     xlab = "Massa Corporal (g)", ylab = "Comprimento do Bico (mm)")

# Reta Original (Item 2) em Azul
abline(modelo_linear, col = "blue", lwd = 2)

# Reta com Outlier (Item 4) em Vermelho Tracejado
abline(modelo_outlier, col = "red", lwd = 2, lty = 2)

# Legenda Tecnica
legend("topleft", legend = c("Modelo Original", "Modelo com Outlier"),
       col = c("blue", "red"), lty = c(1, 2), lwd = 2, bty = "n")
grid(col = "lightgray", lty = "dotted")
```