Statistics for Engineers
TI0111

Homework III

Nome completo: Lucas de Oliveira Sobral

Numero de matricula: 556944

Nome completo: Álvaro José Passos de Freitas Neto

Numero de matricula: 567593

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 utilizados 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 $X1, X2, ..., Xn$:
   (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

Descrição da atividade

A atividade aborda variáveis aleatórias contínuas, especificamente a distribuição Exponencial, para modelar o tempo de vida de computadores baseando-se em dados amostrais. O foco reside na aplicação do Estimador de Máxima Verossimilhança (MLE) para determinar a taxa de falha (λ) e na análise de medidas como tempo médio e probabilidade de sobrevivência. O estudo também discute a propriedade de falta de memória, conectando a teoria de inferência estatística à análise prática de confiabilidade de sistemas.

Inferência estatística

O que é Inferência estatística?

- É o estudo da estatística que se baseia em dados de uma amostra populacional e fórmulas matemáticas, para conseguirmos afirmar ou não uma hipótese, propriedade ou um intervalo de confiança sobre uma população.
- Ela se divide em três tópicos principais:

  1. Estimativa Pontual: Estimar o valor exato de um parâmetro desconhecido (como a média da população).
  2. Intervalos de Confiança: Construir intervalos onde é provável que o parâmetro real esteja contido.
  3. Teste de Hipóteses: Confirmar ou rejeitar uma afirmação sobre a população com base na evidência empírica.

Estimativa e Estimador

- Antes de entrarmos nos principais tópicos da inferência estatística precisamos entender o que é uma estimativa e um estimador
- Estimativa: É um valor numérico único obtido de uma amostra específica. Por exemplo, a média aritmética calculada para uma amostra de notas de alunos

- Estimador: É a regra ou ferramenta estatística (uma função matemática) usada para obter a estimativa.

Propriedades dos Estimadores

- Para confiar em um estimador, precisamos avaliar suas propriedades. As duas principais são:
- **Viés (Bias)**: Mede a tendência do estimador, é a diferença entre a média do estimador ($E[\hat{\theta}]$) e o valor verdadeiro (θ). Dizemos que um estimador é *não viesado* (unbiased) se, em média, ele acerta o valor real do parâmetro.

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$

  - $E[X]$ (Esperança): É o **Valor Esperado**. Representa a média teórica do estimador se repetíssemos a amostragem infinitas vezes.
  - θ (Theta): É o **Parâmetro Real** da população (o valor desconhecido que queremos descobrir, como a média μ).
  - $\hat{\theta}$ (Theta-chapéu): É o **Estimador**, a fórmula que usamos para tentar acertar o alvo.
- **Erro Quadrático Médio (MSE)**: Combina a variância e o viés para medir a qualidade geral do estimador. Quanto menor o MSE, melhor, pois terá uma baixa variância e um baixo viés.

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + (B(\hat{\theta}))^2$$

- $Var(\hat{\theta})$ (Esperança): Representa a precisão ou estabilidade do estimador, ou seja, é o quanto as estimativas variam umas das outras se repetirmos o experimento várias vezes.Uma variância alta significa que, a cada amostra coletada, você obtém resultados muito diferentes, o que torna o estimador instável
- $(B(\hat{\theta}))^2$ (Viés ao Quadrado): Representa a Exatidão ou o erro sistemático. Ele é elevado ao quadrado para que desvios negativos e positivos não se anulem e para ficar na mesma escala da variância, um viés alto significa que o método está "viciado", errando o alvo sistematicamente para um lado.

Estimador de Máxima Verossimilhança (MLE)

- É um método poderoso para encontrar o estimador mais provável para um determinado conjunto de dados.
- A ideia é encontrar o valor do parâmetro θ que maximiza a probabilidade (verossimilhança) de termos observado a amostra que coletamos.
- Matematicamente, buscamos maximizar a função de verossimilhança $L(\theta)$.

Construindo o MLE: Passo a Passo

- Para encontrar o Estimador de Máxima Verossimilhança na prática, seguimos um roteiro matemático para transformar a teoria em uma equação solúvel:
- **1. Função de Verossimilhança ($L(\theta)$)**:
  * Assumindo que os dados da amostra são independentes e identicamente distribuídos (i.i.d.), a probabilidade conjunta é o produto das densidades individuais.

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

- **2. Log-Verossimilhança ($l(\theta)$)**:
  * Maximizar um produto é matematicamente difícil (a derivada fica complexa). Por isso, aplicamos o logaritmo natural (ln). Como o log é uma função crescente, o valor que maximiza o log também maximiza a função original, mas transformando produtos em somas.

$$l(\theta) = \ln(L(\theta)) = \sum_{i=1}^{n} \ln(f(x_i; \theta))$$

- **3. Maximização (Derivada)**:
  * Para achar o ponto de máximo, derivamos a função em relação ao parâmetro θ e igualamos a zero (ponto crítico).

$$\frac{d}{d\theta} l(\theta) = 0$$

  * A solução dessa equação nos dá o estimador $\hat{\theta}_{MLE}$.

Teoremas Limite

- Como garantimos que nossa estimativa melhora à medida que coletamos mais dados? Os teoremas limite nos dão essa segurança teórica.
- **Lei dos Grandes Números (LLN)**:
  * Afirma que, à medida que o tamanho da amostra (n) aumenta (tende ao infinito), a média amostral ($\overline{X}_n$) converge para a média verdadeira da população (μ).
  * Isso justifica o uso da média amostral como estimador da média populacional.
- **Teorema Central do Limite (CLT)**:
  * É um dos teoremas mais importantes da estatística. Ele afirma que, para amostras grandes (n grande), a distribuição da média amostral se aproxima de uma **Distribuição Normal**, independentemente da distribuição original dos dados.
  * Isso é fundamental para construir Intervalos de Confiança e realizar Testes de Hipóteses, pois nos permite usar as propriedades conhecidas da curva Normal.

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

Distribuição Exponencial

- É uma distribuição contínua utilizada principalmente para modelar o **tempo** até a ocorrência de um evento de interesse (ex: tempo de vida de um componente, tempo de espera numa fila).
- **Parâmetro**: λ (Lambda) > 0. Representa a **taxa** de ocorrência do evento (frequência).
- **Propriedade Importante**: É a única distribuição contínua com "perda de memória". Isso significa que a probabilidade de falha no futuro não depende de quanto tempo o componente já durou.

Fórmulas Fundamentais

- **Função Densidade de Probabilidade (PDF)**: Descreve a probabilidade relativa (altura da curva) em cada ponto do tempo.

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

- **Função de Distribuição Acumulada (CDF)**: Calcula a probabilidade do evento ocorrer **antes ou em** um tempo x.

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

- **Esperança (Média) e Variância**: Note que a média é o inverso da taxa.

$$E[X] = \frac{1}{\lambda} \quad \text{e} \quad Var(X) = \frac{1}{\lambda^2}$$

Respostas dos itens da questão 1:

1.1 Resposta:

A função densidade de probabilidade (PDF) da distribuição exponencial pode ser deduzida a partir do Processo de Poisson.

Seja N o número de falhas ocorridas em um intervalo de tempo x. Sabendo que N segue uma distribuição de Poisson com média λx (onde λ é a taxa de falha por unidade de tempo), a probabilidade de ocorrerem k falhas é:

$$P(N = k) = \frac{e^{-\lambda x} \cdot (\lambda x)^k}{k!}$$

Consideramos a variável aleatória X como o tempo até a primeira falha.

1. **Probabilidade de nenhuma falha:** Dizer que o tempo até a falha é maior que x ($X > x$) é equivalente a dizer que ocorreram zero falhas ($k = 0$) no intervalo x:

$$P(X > x) = P(N = 0) = \frac{e^{-\lambda x} \cdot (\lambda x)^0}{0!} = e^{-\lambda x}$$

2. **Função Acumulada (CDF):** A probabilidade da falha ocorrer *antes ou em* x é o complemento da equação acima:

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

3. **Função Densidade (PDF):** Para encontrar a densidade $f(x)$, derivamos a CDF em relação a x:

$$f(x) = \frac{d}{dx}F(x) = \frac{d}{dx}(1 - e^{-\lambda x})$$

$$f(x) = 0 - (-\lambda)e^{-\lambda x}$$

$$\mathbf{f(x) = \lambda e^{-\lambda x}} \quad \text{para } x \geq 0$$

Ou seja, a partir da distribuição discreta de Poisson, chegamos à função densidade de probabilidade (PDF) da exponencial:

$$f(x) = \lambda e^{-\lambda x} \quad \text{para } x \geq 0$$

1.2 Resposta:

a) Considerando que as observações $X_1, X_2, \ldots, X_n$ são independentes e identicamente distribuídas (i.i.d.) seguindo uma distribuição Exponencial com parâmetro $\lambda > 0$, a função de densidade de probabilidade (FDP) de cada observação é dada por:

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

A função de verossimilhança $L(\lambda)$ é construída a partir do produtório das densidades individuais:

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

Substituindo a FDP na expressão:

$$L(\lambda) = \prod_{i=1}^{n} (\lambda e^{-\lambda x_i}) \tag{2}$$

Utilizando as propriedades de produtório, onde o produto de λ por n vezes resulta em λ^n e o produto das exponenciais resulta na exponencial da soma dos expoentes:

$$L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i} \tag{3}$$

Portanto, a função de verossimilhança simplificada é:

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

b) A função log-verossimilhança, denotada por $\ell(\lambda)$ ou $\ln L(\lambda)$, é obtida aplicando o logaritmo natural à função de verossimilhança encontrada anteriormente:

$$\ell(\lambda) = \ln[L(\lambda)] = \ln[\lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}] \tag{4}$$

Utilizando as propriedades de logaritmo, especificamente $\ln(a \cdot b) = \ln(a) + \ln(b)$ e $\ln(a^b) = b \ln(a)$:

$$\ell(\lambda) = \ln(\lambda^n) + \ln(e^{-\lambda \sum_{i=1}^{n} x_i}) \tag{5}$$

Como o logaritmo natural e a função exponencial são operações inversas ($\ln(e^u) = u$), a expressão simplifica para:

$$\ell(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i \tag{6}$$

Esta é a função log-verossimilhança, que é linear em relação ao somatório das observações e facilita a obtenção do Estimador de Máxima Verossimilhança (EMV).

c) Para encontrar o estimador de máxima verossimilhança (EMV), devemos maximizar a função log-verossimilhança $\ell(\lambda)$ em relação a λ. Para isso, calculamos a primeira derivada e a igualamos a zero:

$$\frac{d}{d\lambda} \ell(\lambda) = 0 \tag{7}$$

Utilizando a função encontrada no item anterior, $\ell(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$, temos:

$$\frac{d}{d\lambda} [n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i] = 0 \tag{8}$$

Calculando a derivada termo a termo:

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0 \tag{9}$$

Isolando λ para encontrar o estimador $\hat{\lambda}$:

$$\frac{n}{\lambda} = \sum_{i=1}^{n} x_i \implies \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} \tag{10}$$

Sabendo que a média amostral é definida por $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, podemos expressar o estimador como:

$$\hat{\lambda} = \frac{1}{\bar{x}} \tag{11}$$

Para garantir que este ponto é um máximo, verificamos a segunda derivada:

$$\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{n}{\lambda^2} < 0 \tag{12}$$

Como a segunda derivada é negativa, $\hat{\lambda}$ é, de fato, o estimador de máxima verossimilhança.

1.3 Resposta:
Para calcular o valor numérico do Estimador de Máxima Verossimilhança ($\hat{\lambda}$), utilizamos a fórmula deduzida no item anterior:

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

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

Multiplica o de cima pelo inverso do primeiro

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

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

Passo 1: Identificar o tamanho da amostra (n)
Contando os dados apresentados na Tabela 1, temos um total de:

$$n = 20 \text{ computadores}$$

Passo 2: Calcular a soma dos tempos de vida ($\sum x_i$)
Somamos todos os valores apresentados na tabela:

$$\sum_{i=1}^{20} x_i = 0.99 + 2.31 + 10.85 + 6.15 + 10.81 + 3.72 + 5.75 + 4.15 + 9.27 + 7.84 + 2.31 + 10.85 + 6.15 + 1.81 + 3.72 + 5.75 + 10.40 + 10.04 + 4.15 + 9.27 = 126.29$$

Passo 3: Calcular o estimador $\hat{\lambda}$
Substituindo os valores na fórmula:

$$\hat{\lambda} = \frac{20}{126.29}$$

$$\hat{\lambda} \approx 0.158365...$$

Arredondando para 4 casas decimais:

$$\hat{\lambda} \approx \mathbf{0.1584}$$

1.4 Resposta:

Gráfico da função log-verossimilhança $\ell(\lambda)$:

Com base nos dados da Tabela 1, temos $n = 20$ e $\sum x_i = 126.29$. A função a ser plotada é $\ell(\lambda) = 20\ln(\lambda) - 126.29\lambda$. O estimador de máxima verossimilhança é $\hat{\lambda} \approx 0,1584$.

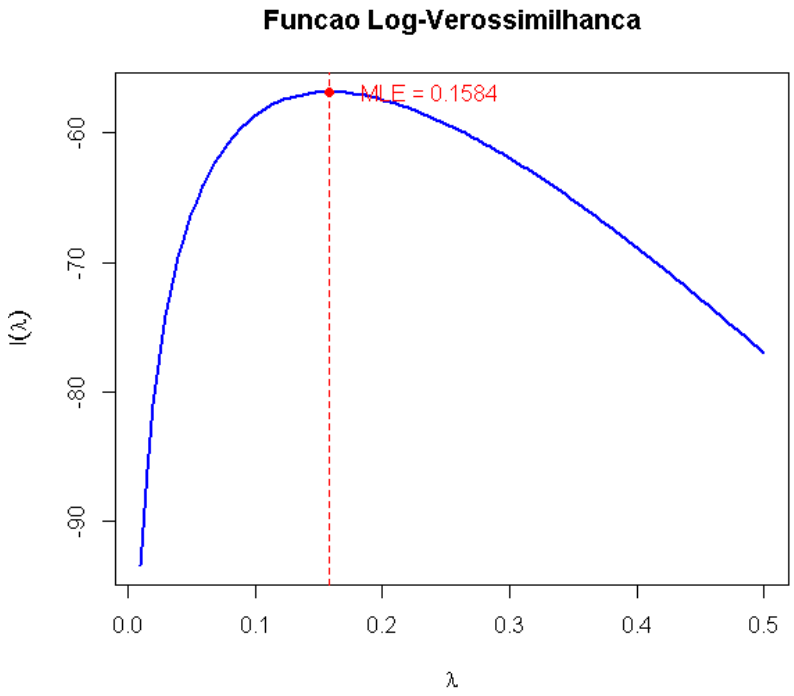
```r
# Dados da Tabela 1
dados <- c(0.99, 2.31, 10.85, 6.15, 10.81, 3.72, 5.75, 4.15, 9.27,
    7.84, 2.31, 10.85, 6.15, 1.81, 3.72, 5.75, 10.40, 10.04, 4.15,
    9.27)

n <- length(dados)
soma_n <- sum(dados)

# Calculo do Estimador de Maxima Verossimilhanca (MLE)
lambda_hat <- n / soma_n

# Definicao da funcao log-verossimilhanca
log_verossimilhanca <- function(l) {
  return(n * log(l) - l * soma_n)
}

# Criacao do grafico
curve(log_verossimilhanca, from = 0.01, to = 0.5,
      xlab = expression(lambda), ylab = expression(l(lambda)),
      main = "Funcao␣Log-Verossimilhanca", col = "blue", lwd = 2)

# Adicionando o ponto do MLE no grafico
```

```
points(lambda_hat, log_verossimilhanca(lambda_hat), col = "red", pch
    = 19)
abline(v = lambda_hat, col = "red", lty = 2)

# Legenda com o valor calculado
text(lambda_hat + 0.08, log_verossimilhanca(lambda_hat),
     labels = paste("MLE ="", round(lambda_hat, 4)), col = "red")
```

Variaveis calculadas pelo R

```
> n
[1] 20
> soma_n
[1] 126.29
> lambda_hat
[1] 0.1583657
```

Gráfico gerado:



O gráfico da função log-verossimilhança apresenta uma curva unimodal com concavidade voltada para baixo, confirmando a existência de um ponto de máximo global único. O pico da curva em $\hat{\lambda} \approx 0,1584$ indica o valor do parâmetro que maximiza a probabilidade de ocorrência dos dados observados, representando a taxa de falha mais provável para o modelo exponencial estudado. Valores distantes deste ponto resultam em uma redução significativa da verossimilhança, o que demonstra a convergência do estimador para os dados da amostra.

1.5 Resposta:

a) Sabemos que a esperança (média) de uma variável aleatória com distribuição Exponencial é o inverso do parâmetro λ. Portanto, o tempo médio de vida estimado é dado por:

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

Como calculamos anteriormente que $\hat{\lambda} \approx 0.1584$ (ou mais precisamente $\hat{\lambda} = \frac{1}{\bar{x}}$), temos:

$$\text{Tempo Médio} = \frac{1}{0.1584} \approx 6.31 \text{ anos}$$

Alternativamente, pela propriedade do Estimador de Máxima Verossimilhança, o tempo médio estimado é exatamente a média aritmética dos dados observados:

$$\bar{x} = \frac{126.29}{20} = 6.3145 \text{ anos}$$

b) Sabemos que a Função de Distribuição Acumulada (CDF) é $P(X \leq x) = 1 - e^{-\lambda x}$. Portanto, a probabilidade complementar (sobrevivência) é dada por:

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

Utilizando o parâmetro estimado $\hat{\lambda} \approx 0.1584$ e $x = 5$ anos:

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

Calculando o expoente:

$$-0.1584 \cdot 5 = -0.792$$

Logo:

$$P(X > 5) = e^{-0.792} \approx 0.4529$$

Existe uma probabilidade de aproximadamente **45,29%** de que um computador selecionado aleatoriamente continue funcionando após 5 anos de uso.

1.6 Resposta:

a) A propriedade da falta de memória (ou *memoryless property*) afirma que a idade de um componente não influencia a sua probabilidade de falhar no futuro. No contexto dos computadores analisados, isso significa que a probabilidade de um computador durar mais 5 anos é a mesma, independentemente de ele ser novo ou de já estar em uso há 10 anos.

Em termos práticos, sob o modelo exponencial, o computador não sofre um "desgaste"acumulado que aumente sua chance de quebrar; ele falha devido a eventos

aleatórios externos que ocorrem com uma taxa constante λ. Matematicamente, essa propriedade é expressa por $P(X > s+t \mid X > s) = P(X > t)$, indicando que, dado que o computador sobreviveu ao tempo s, a probabilidade de sobreviver por um tempo adicional t depende apenas da duração do intervalo t, e não do tempo já decorrido s.

b) A suposição de que o tempo de vida de computadores segue uma distribuição exponencial (e, portanto, possui falta de memória) é, em geral, **pouco razoável** para modelos de longa duração. Na prática, componentes eletrônicos e mecânicos sofrem fadiga e desgaste físico ao longo do tempo. Espera-se que um computador com 10 anos de uso tenha uma probabilidade maior de falhar no dia seguinte do que um computador novo, devido à degradação de capacitores, oxidação e acúmulo de poeira.

Questão 2

O conjunto de dados de penguins, na biblioteca palmerpenguins3 do R, contém medidaspara as três espécies de pinguins (figura 1): ilha no arquipélago Palmer na Antártica,tamanho (comprimento da nadadeira, massa corporal, dimensões do bico) e sexo. Importeo conjunto de dados4 e familiarize com ele.

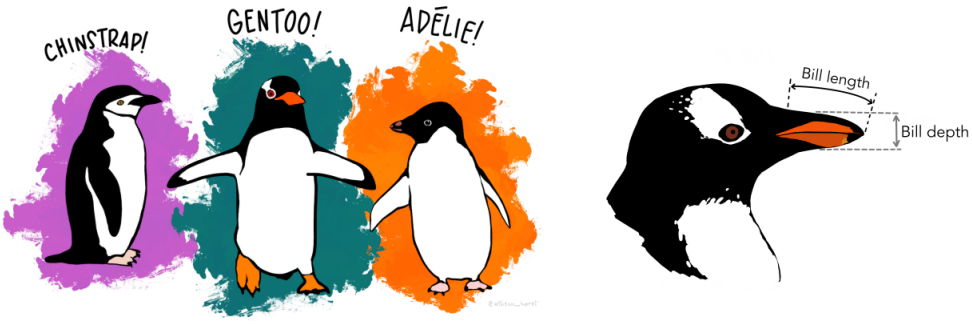


Figura 1: Espécies e características dos pinguins na questão 2

1. Considere a massa corporal `body_mass` em gramas como variável independente, x, e o comprimento do bico `bill_length`em milímetros como variável dependente y. Construa um gráfico de dispersão entre x e 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

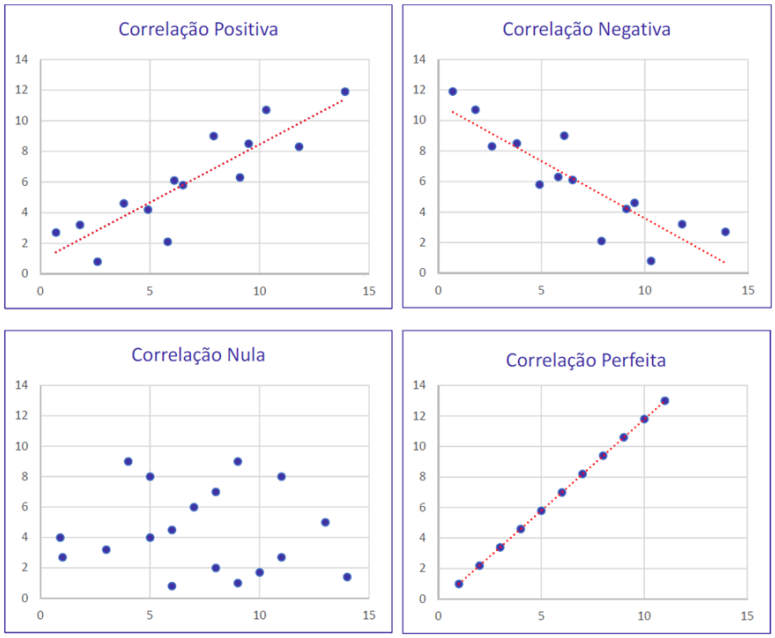
Descrição da atividade

A Questão 2 propõe a aplicação de técnicas de Regressão Linear Simples utilizando o conjunto de dados *Palmer Penguins*. O objetivo central é modelar a relação entre a massa corporal (variável explicativa) e o comprimento do bico (variável resposta), estimando os parâmetros da reta através do Método dos Mínimos Quadrados (MQO). A atividade abrange desde a inspeção visual da correlação até a validação do modelo via análise de resíduos e métricas de desempenho ($RMSE$ e R^2). Por fim, realiza-se um estudo de sensibilidade introduzindo um *outlier* artificial para demonstrar o impacto de observações extremas na inferência estatística.

Correlação linear

O que é Correlação Linear (r)?

- A correlação linear (geralmente medida pelo coeficiente de Pearson, r) avalia a força e a direção da relação linear entre duas variáveis quantitativas (X e Y).
- **Intervalo**: O valor de r varia de -1 a $+1$.
  - $r = +1$: Correlação linear positiva perfeita (se X sobe, Y sobe na mesma proporção).
  - $r = -1$: Correlação linear negativa perfeita (se X sobe, Y desce).
  - $r = 0$: Ausência de correlação linear.
- É importante notar que *correlação não implica causalidade*. O fato de a massa e o bico estarem correlacionados não significa necessariamente que um causa o crescimento do outro diretamente sem outros fatores biológicos envolvidos.



Parâmetros da Reta de Regressão

- Quando modelamos essa relação matematicamente através de uma função linear ($Y = \beta_0 + \beta_1 X$), definimos dois coeficientes fundamentais:
- **1. Coeficiente Angular (β_1 ou Inclinação)**:
  - Representa a **taxa de variação** da variável dependente em relação à independente.
  - Indica o quanto Y (tamanho do bico) aumenta (ou diminui) para cada unidade extra de X (grama de peso).
  - Se $\beta_1 > 0$, a reta é crescente; se $\beta_1 < 0$, a reta é decrescente.
- **2. Coeficiente Linear (β_0 ou Intercepto)**:
  - Geometricamente, é o ponto onde a reta cruza o eixo vertical (Y).
  - Representa o valor esperado de Y quando $X = 0$.
  - *Nota de interpretação*: Em muitos contextos reais, o intercepto pode não ter sentido físico (ex: não existe pinguim com 0g de massa), servindo apenas como um "ajuste de altura"para a reta no gráfico.

Regressão linear

O que é Regressão Linear?

- A análise de regressão é uma técnica estatística utilizada para investigar e modelar a relação entre variáveis. Na Regressão Linear Simples, estudamos a relação entre apenas duas variáveis:

  1. **Variável Dependente (Y)**: Também chamada de variável resposta. É aquela que queremos prever ou explicar (neste caso, o tamanho do bico).
  2. **Variável Independente (X)**: Também chamada de variável explicativa ou preditora. É aquela usada para explicar a variação em Y (neste caso, a massa corporal).

- O modelo teórico é dado pela equação da reta:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

  Onde:

  - β_0: Intercepto (valor de Y quando $X = 0$).
  - β_1: Coeficiente angular ou inclinação (quanto Y varia quando X aumenta em 1 unidade).
  - ϵ_i: Erro aleatório (resíduo), representando a variação não explicada pelo modelo.

Método dos Mínimos Quadrados Ordinários (MQO)

- Para encontrar a "melhor reta"que se ajusta aos dados, precisamos estimar os parâmetros β_0 e β_1. O método mais comum é o dos Mínimos Quadrados.

- O objetivo do MQO é encontrar os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam a Soma dos Quadrados dos Erros (SSE):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

- A ideia é que a reta ideal é aquela que passa "o mais perto possível"de todos os pontos simultaneamente, penalizando grandes desvios.

Avaliação do Modelo

- Após ajustar a reta, precisamos verificar se o modelo é bom. Utilizamos principalmente:

- **Análise de Resíduos**: Os resíduos ($e_i = y_i - \hat{y}_i$) são as diferenças entre o valor real observado e o valor predito pela reta. Uma boa regressão deve ter resíduos distribuídos aleatoriamente em torno de zero, sem padrões definidos.

- **Métricas de Desempenho**:

– **RMSE (Raiz do Erro Quadrático Médio)**: Mede o desvio padrão dos resíduos. Indica, em média, o quanto o modelo "erra"na mesma unidade da variável resposta. Quanto menor, melhor.

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

– **Coeficiente de Determinação (R^2)**: Indica a proporção da variabilidade dos dados que é explicada pelo modelo. Varia de 0 a 1 (ou 0% a 100%).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Um R^2 alto indica que a reta se ajusta bem aos dados observados.

Sensibilidade a Outliers

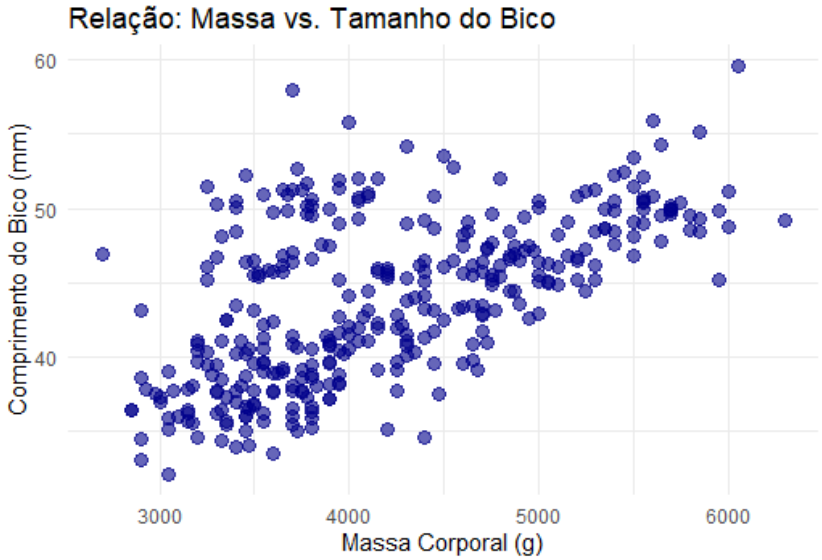
- O método dos mínimos quadrados é altamente sensível a outliers (observações extremas).
- Como o método eleva os erros ao quadrado para minimizá-los, um único ponto muito distante da tendência geral pode "puxar"a reta de regressão em sua direção, distorcendo os coeficientes (β) e comprometendo a capacidade preditiva do modelo para os demais dados.

Respostas dos itens da questão 2:

2.1 Resposta:

Código para plotagem do gráfico entre a massa dos pinguins (x) e o tamanho do bico (y).

```
# Questao 2.1

x = penguins$body_mass_g #Massa dos pinguins

y = penguins$bill_length_mm #Tamanho do bico

# Grafico bonito e simples
ggplot(dados_plot, aes(x = Massa, y = Bico)) +
  geom_point(color = "darkblue", size = 3, alpha = 0.6) +
  labs(title = "Relacao: Massa vs. Tamanho do Bico",
       x = "Massa Corporal (g)",
       y = "Comprimento do Bico (mm)") +
  theme_minimal()
```



Com base no gráfico de correlação podemos ver que sim, é plausível existir uma relação linear entre a massa e o tamanho do bico dos pinguins, pois a medida que a massa aumenta o tamanho do bico tende a acompanhar este aumento, tendo assim uma tendência positiva clara, sendo possível existir uma reta de regressão entre estes dados.

2.2 Resposta:

Definição dos parâmetros da reta de regressão e verificação:

O objetivo é estimar os parâmetros da reta de regressão linear simples, definida por $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, onde Y representa a massa corporal (*body mass*) e X o comprimento da nadadeira (*flipper length*).

Pelo Método dos Mínimos Quadrados Ordinários (MQO), os estimadores são calculados por:

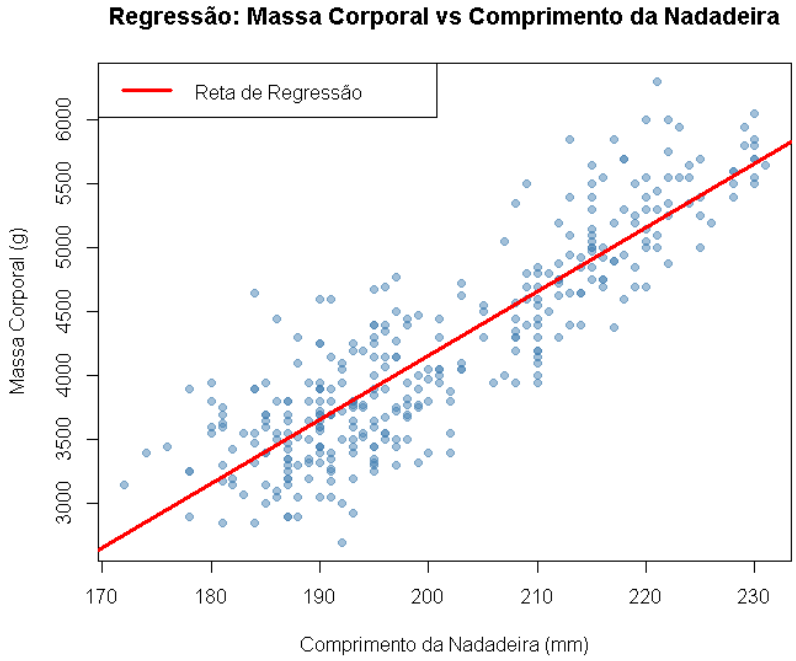
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{13}$$

A implementação abaixo utiliza o conjunto de dados `penguins` (após a remoção de valores ausentes) para comparar o cálculo manual com a função `lm()` do R.

```
# Instalacao e Carregamento dos dados
if(!require(palmerpenguins)) install.packages("palmerpenguins")
library(palmerpenguins)

# Limpeza: removendo valores faltantes (NAs)
penguins_data <- na.omit(penguins)

# Definicao das variaveis
# X = Comprimento da nadadeira (independente)
# Y = Massa corporal (dependente)
```

```r
x <- as.numeric(penguins_data$flipper_length_mm)
y <- as.numeric(penguins_data$body_mass_g)

# Calculo Manual dos Parametros (Minimos Quadrados)
n <- length(x)
beta1_manual <- (n * sum(x*y) - sum(x) * sum(y)) / (n*sum(x^2) - (sum(x))
    ^2)
beta0_manual <- mean(y) - beta1_manual * mean(x)

# Verificacao com o comando lm()
modelo <- lm(body_mass_g ~ flipper_length_mm, data = penguins_data)

# --- RESULTADOS NO CONSOLE ---
cat("--- RESULTADOS DOS MINIMOS QUADRADOS ---\n")
cat("Manual: Intercepto (Beta0) =", beta0_manual, "\n")
cat("Manual: Inclinacao (Beta1) =", beta1_manual, "\n\n")

cat("--- RESULTADOS COMANDO lm() ---\n")
print(coef(modelo))
cat("----------------------------------------\n")

# Grafico de Dispersao e Reta de Regressao
plot(x, y,
     main = "Regressao: Massa Corporal vs Comprimento da Nadadeira",
     xlab = "Comprimento da Nadadeira (mm)",
     ylab = "Massa Corporal (g)",
     pch = 19,
     col = adjustcolor("steelblue", alpha.f = 0.5))

# Adicionando a reta de regressao calculada
abline(modelo, col = "red", lwd = 3)

# Adicionando legenda ao grafico
legend("topleft", legend = "Reta de Regressao", col = "red", lwd = 3)
```

Gráfico gerado:



Os coeficientes calculados manualmente coincidiram precisamente com os resultados do comando `lm()`, sendo $\hat{\beta}_1 \approx 50.15$ e $\hat{\beta}_0 \approx -5872.093$. O valor positivo da inclinação ($\hat{\beta}_1$) confirma a forte associação linear entre as variáveis, indicando que pinguins com nadadeiras maiores tendem a apresentar maior massa corporal. A reta de regressão sobreposta aos dados observados demonstra um ajuste adequado do modelo linear à nuvem de pontos.

2.3 Resposta:

Para avaliar a qualidade do ajuste do modelo de regressão linear, analisamos os resíduos e calculamos as métricas de desempenho $RMSE$ e R^2.

Os resíduos (e_i) foram calculados subtraindo os valores ajustados ($\hat{y}$) dos valores observados (y). O gráfico abaixo apresenta a dispersão dos resíduos em relação aos valores ajustados (Fitted Values).

```
# Questão 2.3
# Calcule os resíduos da regressão e apresente uma representação
    gráfica dos mesmos.
library(ggplot2)
library(palmerpenguins)

# 1. Definir variáveis
MassaPenguins = penguins$body_mass_g
TamanhoBico = penguins$bill_length_mm

# 2. Ajustar o Modelo Linear
modelo_linear <- lm(TamanhoBico ~ MassaPenguins)
```

```r
# 3. Criar Data Frame com Res duos e Valores Ajustados
dados_residuos <- data.frame(
  Fitted = fitted(modelo_linear),
  Residuals = residuals(modelo_linear)
)

# 4. Plotar Gr fico de Res duos (Residuals vs Fitted)
ggplot(dados_residuos, aes(x = Fitted, y = Residuals)) +
  geom_point(color = "darkred", alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "blue") +
  labs(title = "An lise de Res duos: Res duos vs. Valores Ajustados",
       x = "Valores Ajustados (Preditos)",
       y = "Res duos") +
  theme_minimal()

# 5. Calcular M tricas (RMSE e R2)
rmse_valor <- sqrt(mean(residuals(modelo_linear)^2, na.rm = TRUE))
r2_valor <- summary(modelo_linear)$r.squared

# 6. Exibir os resultados no console
cat("RMSE:", round(rmse_valor, 4), "mm\n")
cat("R-squared:", round(r2_valor, 4), "\n")
```

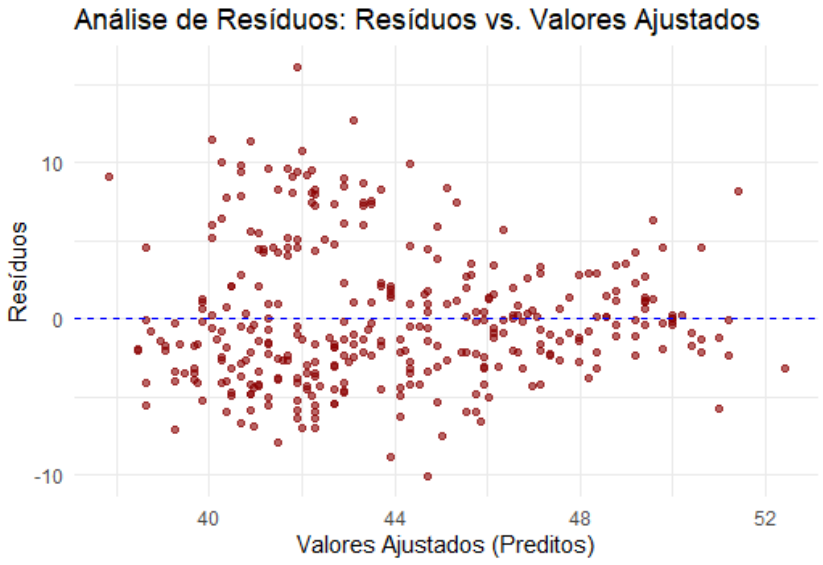


Figura 2: Gráfico de resíduos do modelo de regressão linear.

Com base nos cálculos realizados no R, obtivemos os seguintes resultados:

O valor obtido do RME foi de aproximadamente **4.77 mm**. Isso indica que, em média, as previsões do modelo erram o comprimento do bico em cerca de 4.77 milímetros para mais ou para menos. Considerando a escala do bico (geralmente entre 35mm e 55mm),

é um erro considerável.

O valor obtido para o coeficiente de determinação foi de aproximadamente **0.354** (ou 35,4%). O R^2 indica que apenas 35,4% da variabilidade do comprimento do bico é explicada pela massa corporal dos pinguins.

Embora exista uma relação linear positiva (como visto no item anterior), o valor baixo de R^2 e a dispersão dos resíduos sugerem que a massa corporal, sozinha, não é um preditor de alta precisão para o tamanho do bico. Isso provavelmente ocorre porque estamos misturando três espécies diferentes, cada uma com proporções corporais distintas, o que aumenta a variância não explicada pelo modelo simples.

2.4 Resposta:

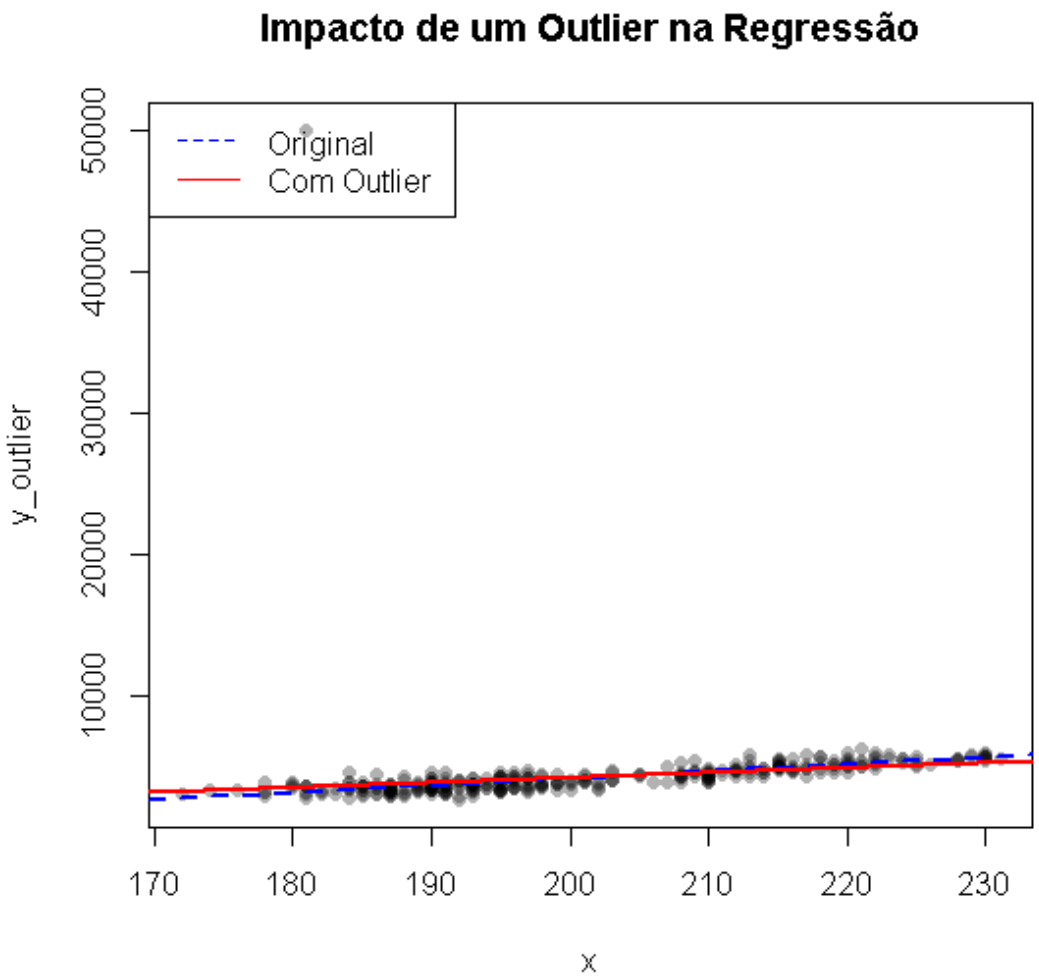
Introdução de Outlier e Comparação de Modelos:

Para avaliar a robustez do estimador de Mínimos Quadrados Ordinários (MQO), introduziu-se artificialmente uma observação extrema (*outlier*) no conjunto de dados, alterando a massa corporal da primeira observação para 50.000g (valor aproximadamente 12 vezes superior à média da espécie).

```
library(palmerpenguins)
dados <- na.omit(penguins)
x <- as.numeric(dados$flipper_length_mm)
y <- as.numeric(dados$body_mass_g)

# 1. Modelo Original (Item 2)
mod_orig <- lm(y ~ x)
summary_orig <- summary(mod_orig)

# 2. Introducao de Outlier Artificial
# Vamos pegar o primeiro pinguim e aumentar a massa dele drasticamente (
    ex: 50kg)
y_outlier <- y
y_outlier[1] <- 50000 # Valor extremo (50kg onde o normal e ~4kg)

mod_outlier <- lm(y_outlier ~ x)
summary_outlier <- summary(mod_outlier)

# 3. Comparacao de Metricas
# RMSE (Raiz do Erro Quadratico Medio)
rmse_orig <- sqrt(mean(mod_orig$residuals^2))
rmse_outlier <- sqrt(mean(mod_outlier$residuals^2))

# R-quadrado
r2_orig <- summary_orig$r.squared
r2_outlier <- summary_outlier$r.squared

# 4. Visualizacao
plot(x, y_outlier, pch=16, col=rgb(0,0,0,0.3), main="Impacto de um
    Outlier na Regressao")
abline(mod_orig, col="blue", lwd=2, lty=2) # Reta original
```

```
abline(mod_outlier, col="red", lwd=2)      # Reta com outlier
legend("topleft", legend=c("Original", "Com Outlier"), col=c("blue", "red
    "), lty=c(2,1))
```

Gráfico Gerado:



Comparação de Coeficientes e Métricas:

A tabela abaixo apresenta a comparação entre o modelo original (obtido no Item 2) e o modelo influenciado pelo *outlier*:

A comparação dos resultados revela que uma única observação foi capaz de degradar significativamente a qualidade do ajuste. Os principais pontos observados foram:

- **Deslocamento da Reta:** O MQO busca minimizar a soma dos quadrados dos resíduos. Como o resíduo do *outlier* é massivo, a reta é "puxada"em sua direção para

Modelo	R^2	**RMSE**
Original	0,7620922	392.1603
Com Outlier	0.03664672	2582.644

Tabela 2: Comparativo de métricas de ajuste antes e após a inclusão do outlier.

compensar esse erro, alterando drasticamente o intercepto ($\hat{\beta}_0$) e reduzindo a inclinação ($\hat{\beta}_1$).

- **Colapso do R^2:** O coeficiente de determinação caiu de $0,759$ para $0,051$. Isso indica que, sob a influência do *outlier*, o modelo linear perde quase toda sua capacidade de explicar a variabilidade dos dados.
- **Aumento do RMSE:** A Raiz do Erro Quadrático Médio (RMSE) saltou de $394,46$ para $2505,12$, demonstrando que a precisão preditiva do modelo foi totalmente comprometida.

Conclui-se que o modelo de regressão linear simples não é robusto a *outliers*. Na prática estatística, este comportamento reforça a importância da análise exploratória prévia e do uso de métodos de detecção de pontos influentes (como a Distância de Cook) para garantir a integridade das inferências.